DOCUMENT RESUME

ED 442 830 TM 031 250

AUTHOR	Kelkar, Vinaya; Wightman, Linda F.; Luecht, Richard M.
TITLE	Evaluation of the IRT Parameter Invariance Property for the MCAT.
PUB DATE	2000-04-25
NOTE	66p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000). The Graduate School Research Program is sponsored by the Association of American Medical Colleges--Medical College Admission Test.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Ability; *College Entrance Examinations; *Difficulty Level; Ethnicity; Higher Education; *Item Response Theory; *Racial Differences; Validity
IDENTIFIERS	*Invariance; *Medical College Admission Test; One Parameter Model; Three Parameter Model; Two Parameter Model

ABSTRACT

The purpose of this study was to investigate the viability of the property of parameter invariance for the one-parameter (1P), two-parameter (2P), and three-parameter (3P) item response theory (IRT) models for the Medical College Admissions Tests (MCAT). Invariance of item parameters across different gender, ethnic, and language groups and the invariance of ability parameters with respect to test difficulty were assessed. The study also sought to test the stability of ability estimates obtained for random sample "X" using calibrations from different groups and to identify the most efficient IRT model for the MCAT data. All data were from the fall 1994 administration of the MCAT, with 9 random samples of 1,100 drawn from each of 3 test sections (out of 16,520 eligible test takers). The assumption of unidimensionality was first tested, and evidence was found of two or more underlying dimensions affecting test performance. The three IRT models were fit to the dichotomous response data from every sample, and all three showed adequate fit for the MCAT data, with the 1P model item estimates having the smallest estimation error, although the differences among 1P, 2P, and 3P models were very small in magnitude. Evidence was found to support the conclusion that item and ability parameters are stable/invariant with respect to gender and racial/ethnic and language groups for all models. Ability estimates also appear to be invariant with respect to test difficulty for all models. Nine appendixes contain data tables that supplement the discussion. (Contains 28 tables, 59 figures, and 11 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 830

Evaluation of the IRT Parameter Invariance Property for the MCAT

Vinaya Kelkar, Linda F. Wightman, and Richard M. Luecht

University of North Carolina at Greensboro

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

V. Kelkar

Paper presented at the
Annual Meeting of the National Council on Measurement in Education
as part of the symposium entitled
Making An Informed Choice Among IRT Alternatives-
MCAT Graduate Student Research Program
April 25, 2000

TM031250

The Graduate Student Research Program is sponsored by the Association of American Medical Colleges-Medical College Admission Test. For permission to reprint and additional information regarding papers in the *MCAT Monograph Series* contact: MCAT•AAMC •2450 N Street NW •Washington DC 20037.

BEST COPY AVAILABLE

Evaluation of the IRT Parameter Invariance Property for the MCAT

Vinaya Kelkar, Linda F. Wightman, and Richard M. Luecht

University of North Carolina at Greensboro

Introduction

A critical aspect of the Association of American Medical College's (AAMC) evaluation of the viability of computerizing the Medical College Admissions Tests (MCAT) involves selecting an appropriate item response theory (IRT) model. IRT makes it possible to estimate item characteristics, such as item difficulty and discrimination, and examinee proficiencies relative to the same scale. This common scaling capability across items and examinees makes possible technologies such as computer-adaptive testing (CAT). Under CAT, the difficulty of an item can be matched to the proficiency of an examinee and may yield certain efficiencies (e.g. more reliable scores for a fixed test length or constant reliability from shorter tests). IRT also facilitates many technical psychometric procedures such as equating test forms of varied difficulty over time and different examinee samples.

The three most popular IRT models are the one-parameter (1P), two-parameter (2P) and three-parameter (3P) models. Each additional model parameter (i.e., moving from one to two to three parameters) provides greater capability to *fit* the particular idiosyncrasies of an empirical data set containing scored item responses. For example, the 1P model has a single item difficulty parameter, whereas the 3P has an item difficulty parameter, a slope parameter representing the sensitivity of the item to the underlying proficiency or trait, and a third parameter that helps adjust for guessing or other noisy

response behavior on the part of the lower proficiency examinees. The two additional parameters in the 3P model typically allow that model to fit item response data for multiple-choice tests better than the 1P model (see, for example, Lord, 1980). However, obtaining *stable* estimates for the additional parameters in the 2P and 3P IRT models can be complicated. This leads to a commonly encountered trade-off between selecting a better fitting model (like the 2P model or the 3P model) versus selecting a more parsimonious model with robust parameter estimates. For example, many testing organizations involved in certification and licensure testing have decided that estimation stability across varied examinee samples and over time is more important than nominal gains in fit.

This study considers both sides of that trade in the context of the MCAT item banks and for various gender and racial/ethnic groups within the typical MCAT examinee population. The 1P, 2P and 3P are all included in this study to examine the magnitude of improved fit and the degree to which that fit aids in improving the accuracy and stability of examinee scores. This study also looks carefully at the stability of the IRT model parameter estimates for each model across different examinee language, gender and racial/ethnic group samples. This is sometimes referred to as "parameter invariance". That is, the examinees ought to be indifferent to the method of scaling and to which samples are used in the scaling of their response data to calculate the item statistics. In a computerized testing environment, it is unlikely that large, completely random examinee samples of response data can be obtained for all the items.

The study is conducted in two phases. The first phase focuses on the invariance of item parameters with respect to ethnic group, gender group, and English-language group. The second phase focuses on the invariance of ability parameters with respect to test difficulty, since if a CAT for the MCAT were implemented, the difficulty of tests administered to different examinees might vary considerably. This phase also assesses the stability of item parameters estimated from diverse groups of examinees across the three IRT models. The impact of fit and item parameter estimation stability is evaluated relative to the examinees. In a high stakes testing program like the MCAT, the accuracy and stability of the examinees' proficiency scores are the ultimate criteria.

Method

Data

Forms. All data have been drawn from the fall 1994 administration of Form 15 of the MCAT. Analyses and calibrations were conducted separately for each test section: Biological Sciences (BS), Physical Sciences (PS), and Verbal Reasoning (VR).

Samples. Only those test takers who tested at a standard administration of the MCAT, under standard conditions (e.g. standard time limits etc.) and who have no special irregularity flags in their records, were eligible for selection. There are 16,520 eligible test takers. Frequencies and relative frequencies for each ethnic, gender, and language group within the eligible population, are presented in Tables 1a-1c.

Table 1a.

Gender	Frequency	Percent
Female	7671	46.8
Male	8733	53.2

Frequency missing = 116

Table 1b.

Ethnicity	Frequency	Percent
White	8901	57.0
Asian	3852	24.7
Black	1394	8.9
Mexican Am.	520	3.3
Puerto Rican Com.	452	2.9
Puerto Rican (Mainland)	374	2.4
Native Am.	123	0.8

Frequency missing = 904

Table 1c.

Language	Frequency	Percent
English as Primary Language (EPL)	13081	79.2
English as Secondary Language (ESL)	1151	7.0

Frequency missing =2288

Nine random samples of size 1100 were drawn for each test section (BS, PS, and VR). This sample size was chosen to allow adequate stratified random selection within all of the demographic areas of interest.

Two samples (X) and (Y) were selected, without replacement, from among all eligible test takers. Because each sample was selected independently from the total population of eligible test takers, there was some overlap between X and Y samples- BS (74/1100 or 6.7%), PS (83/1100 or 7.5%), VR (23/1100 or 2.1%). Samples X and Y were used to provide baseline data on the amount of random sampling variation (i.e. normal sampling error) to expect in this study.

Gender and racial/ethnic groups with 1100 or more eligible examinees were sampled using stratified random sampling procedures. The samples included males, females, one EPL group, one ESL group, white test takers, black test takers and Asian test takers. The samples and their respective codings are identified in Table 2.

Table 2.

Summary of Test Taker Samples.

	Sample and Subtest Designations		
Sample	BS	PS	VR
Random Sample X	BSX	PSX	VRX
Random Sample Y	BSY	PSY	VRY
Female	BSF	PSF	VRF
Male	BSM	PSM	VRM
English as Primary (EPL)	BSP	PSP	VRP
English as Secondary (ESL)	BSS	PSS	VRS
White	BSW	PSW	VRW
Black	BSB	PSB	VRB
Asian	BSA	PSA	VRA

Since samples X and Y are used to represent the entire test-taker population, it is critical that they should be composed of the same proportion of each gender and racial/ethnic group as represented in the actual population. The proportion of each sampling group within the population and within each random sample X and Y, for all three test sections is presented in Table 3. It suggests that random sample X and random sample Y are representative of the population.

Table 3.

Proportion of each sampling group in the population and within each random sample X and Y.

Group	Population (percent)	Biological Sciences		Physical Sciences		Verbal Reasoning	
		Sample X	Sample Y	Sample X	Sample Y	Sample X	Sample Y
Female	46.8	47.3	46.2	46.3	43.4	46.6	[illegible].7
Male	53.2	52.7	53.8	53.7	56.6	53.4	53.3
Asian	24.7	23.5	23.0	25.7	26.0	23.9	24.3
Black	8.9	8.7	9.7	9.2	9.8	10.3	10.1
White	57.0	56.2	58.0	55.6	56.2	57.2	57.0
EPL	79.2	77.1	79.0	80.3	79.0	78.7	79.7
ESL	7.0	7.5	6.2	6.5	7.1	6.3	5.5

Calibrations and Scalings

Calibrations. Scored data from the nine samples, for all three test sections - Biological Sciences (BS), Physical Sciences (PS), and Verbal Reasoning were calibrated separately to obtain 1P, 2P, and 3P item and person parameter estimates for all 63 items (55 for VR) and 1100 test takers. All calibrations were performed using BILOG 3.11 for windows, (Mislevy & Bock, 1990).

Scalings. Calibrations for different test taker samples were placed on a common IRT scale by using common-item scaling. The scaling used the characteristic-curve method (Stocking and Lord, 1983). Under this approach, estimated "true scores" are equated using least squares. The base scale was set by the calibration of sample X; all other calibrations were scaled to base sample X calibrated scale.

Analyses

Factor Analysis. Unidimensionality is an important assumption for all 3 IRT models under study. Satisfying the assumption of unidimensionality is especially critical

BEST COPY AVAILABLE

in the context of creating item pools for computer adaptive tests. Factor analysis can be used to check the reasonableness of the assumption of unidimensionality with a set of test items (Hambleton & Traub, 1973). Principal factor analyses of tetrachoric correlations were performed on reference sample X within each test section, and for all gender and racial/ethnic group samples in the BS test section. (Merits of using tetrachoric correlations are discussed in McDonald & Ahlawat, 1974). Scree plots of the eigenvalues for test item correlation matrices are useful in identifying whether a dominant first factor is present (Reckase, 1979). These analyses were conducted separately for each of the three MCAT test sections.

Item Ability Regressions Plots (Kingston & Dorans, 1985). Item ability regression plots were completed for each item for each sample to allow for visual inspection of the observed and predicted proportion correct scores at various examinee ability levels. The range of proficiency is divided into a number of intervals (6), and the proportion of people answering the item correctly within each interval is calculated and plotted. The estimated item response function or ICC, derived from the three estimated parameters, is plotted on the same graph. To the extent that the two plots are similar, the model fits the item response data. Plots like these tell us about the distribution of the item parameter estimates and about the appropriateness of the functional form of the ICC model.

Invariance of item parameter estimates.

Average Bias (Residual error of estimation). Bias was evaluated separately for each item parameter for each model across all samples and test sections. The bias statistic for the a-parameter estimate was computed by

$$BIAS_a = \frac{1}{n}\sum_{i=1}^{n} a_{iX} - a_{iZ}$$

where $BIAS_a$ is the bias statistic for the *a*-parameter, a_{iZ} is the *a*-parameter estimate for item I for the focal group (the group being analyzed), a_{iX} is the *a*-parameter estimate for item *i* for the reference group, and n is the number of items. The same procedure was used to compute the bias statistic for *b* and *c* parameter.

Correlational Anaysis. If test data fit the item response model under investigation, there should be a linear relationship between item parameter estimates from the two examinee samples, even if the samples differ in ability, race/ethnicity, or gender (Lord & Novick, 1968). If this linear relationship is not found, it suggests that the item response model does not fit the test data for one or both of the groups. If a model does not fit, item parameter invariance may be challenged. The most critical item parameter is the item difficulty parameter 'b' - if invariance can be established for 'b', invariance for the other parameters usually follows. Scatterplots are used to compare the item difficulty estimates from the various samples. Graphs from two random samples of the same size (sample X and sample Y) provide a baseline for interpreting plots of principal interest (Hambleton & Swaminathan, 1985).

Invariance of ability parameter estimates. Invariance of ability parameter estimates was assessed with respect to population demographic groups and with respect to test difficulty. Estimating ability for examinees in each group, using item calibrations obtained from random sample X, assessed ability parameter invariance with respect to sampling groups. Distributions of these estimates were compared. This was done across

BEST COPY AVAILABLE

models to compare stability across groups and models. Invariance with respect to test difficulty can be established by comparing ability estimates obtained from two or more item sets from the total item pool of interest. The total test was divided into a relatively hard test and an easy test. Ability estimates obtained from these two tests were compared. A strong linear relationship supports a conclusion of invariant ability parameter estimates. Again comparisons are presented in the graphical form. Ability estimates were also computed for two smaller tests, of equivalent difficulty, created from odd and even halves of the total test. The comparison between these estimates serves as baseline for interpreting the hard and easy estimates.

Model Comparison. When comparing between models, there is no single criteria to help choose the model most appropriate for the data. The selection of the model must be based on the amount of evidence gathered in favor of a particular model. All the earlier analyses were conducted for each of the three models under investigation. If the item ability regression plots demonstrated a better fit between the estimated response function and the test data for a particular model, that could serve as one piece of evidence in support of that model. Clearly, a model with the smallest bias in item parameter estimates would be preferred. Invariant item and ability estimates signify a good model/test response data fit. Comparing invariance across models may lead us to favor one model over another. To compare between the 1P model and the more complex 2P and 3P models, a plot of the slope or discrimination parameter 'a', against the item difficulty parameter 'b' is also presented (an a/b plot). Since the 1P model assumes equal

discrimination indices for all its items, the range of the discrimination indices should be small if this assumption is to be viable.

The models were also compared by assessing the stability of the parameter estimates for each model. Ability, 'θ', was estimated for random sample X from item calibrations obtained from the different groups sampled (gender, ethnic, and language). A high correlation between the different ability estimates for sample X indicates stability across groups. Correlations were computed across models to compare stability across groups and models.

To further compare between models, plots of empirical standard errors with 95% error bars, against proficiency (ability) interval midpoints were plotted for each model and for each sample for one test section. These plots were prepared only for the Biological Science test section, but generalize to Physical Science and Verbal Reasoning test sections.

Results

Factor analysis. Scree plots from the factor analysis of the tetrachoric correlations for the test response data, for sample X, for all three test sections (BS, PS, and VR), are presented in Figures 1a-1c. The Scree plot of the Biological Science test section for sample X, in Figure 1a, suggests the presence of a dominant first factor - almost a unidimensional test section - eigenvalue for the first factor is about 7.2 while that for factor 2 is about 2.0. For the Physical Science and Verbal Reasoning test sections, the Scree-plots show the presence of more than one underlying factors.

For the PS test section, in Figure 1b., the eigenvalue for the first factor is about 8.0; almost 3.7 for factor 2 and about 3.1 for factor 3. For the VR section, in Figure 1c.

BEST COPY AVAILABLE

the eigenvalue for factor 1 is about 7.8, while those for factor 2 and 3 are about 3.8 and 3.0 respectively. Although both of these test sections, have one dominant latent factor, both suggest the presence of at least 2-3 additional influential factors.

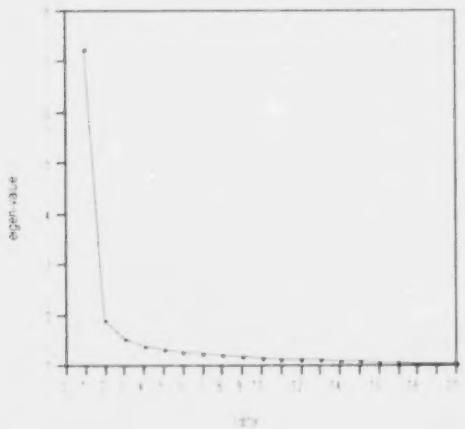


Fig. 1a. Scree plot for Biological Sciences

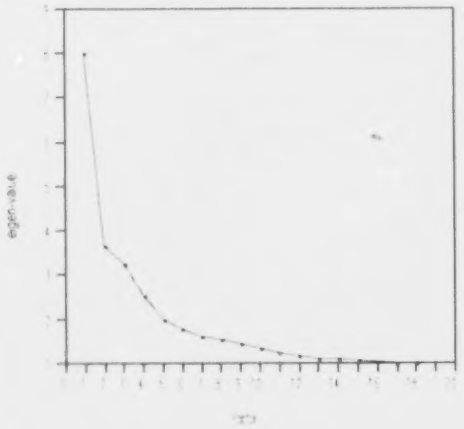


Fig.1b. Scree plot for Physical Sciences.

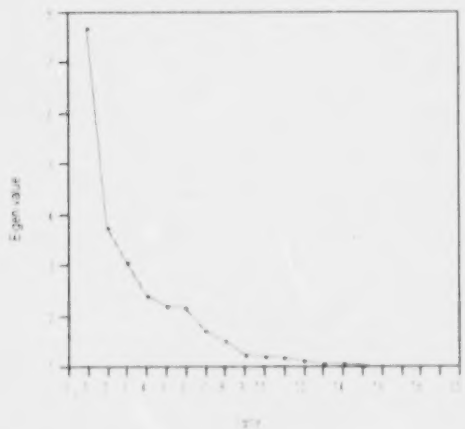


Fig. 1c. Scree plot for Verbal Reasoning.

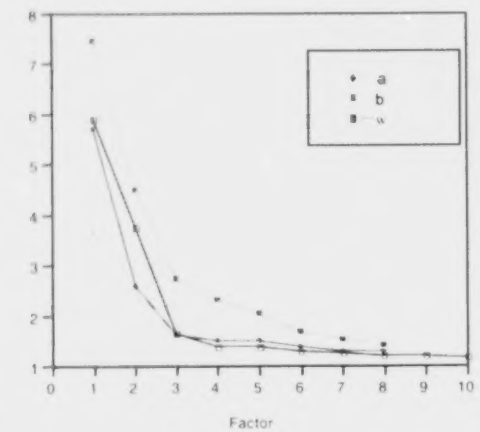


Fig. 1d. Scree plot for ethnic subgroups for BS

To further check the unidimensionality assumption of the BS test section, the responses from the various ethnic groups for Biological Science test section were further factor analyzed. Scree-plots for the ethnic groups are presented in Fig. 1d. For all three ethnic groups: Asian (eigenvalues-5.8, 2.6, 1.7), Black (eigenvalues-7.4, 4.5, 2.8) and White

(eigenvalues-6.0, 3.8, 1.7) the plots suggest the presence of atleast 2-3 significant underlying factors. So, even though the responses for the random sample X suggested a unidimensional BS section, the eigenvalue plots do not indicate unidimensionality for each of the ethnic groups of interest.

Item Ability Regression Plots. To assess the general fit of all 63 items in the Biological Sciences test section, the Physical Sciences section and the 55 items in the Verbal Reasoning section, item ability regression curves were plotted using the 3P model. In addition, to compare fit across models, plots for 2P and 1P for the 63 items in the BS test section were also prepared. To assess fit for the gender and racial/ethnic groups, 3P plots of the 63 items for the Female, Asian and Black groups were completed for the BS section and 3P plots of the 55 items for the ESL group were plotted for the VR section. ESL plots were prepared for the VR test section because this group was hypothesized to be the group most impacted by the subtleties of language found in this section. The entire series of item ability regression plots discussed are presented in Appendix J.

The trace line represents the estimated item response probability function or the Item Characteristic Curve (ICC) for the model. The crosses indicate the observed proportions of correct responses at the various ability points. The size of crosses represents the proportion of people answering the item correctly within each proficiency interval.

Figures 2a - 2h illustrate the fit of IRT model to data from one, representative, item for various gender and racial/ethnic group samples. Figures 2a shows the fit of the 3P model to sample X, while Figures 2b. and 2c. show the 2P and 1P model fit respectively - these are all for the BS test section.

Biological Sciences - item 3

Figure 2a. Sample X - 1P

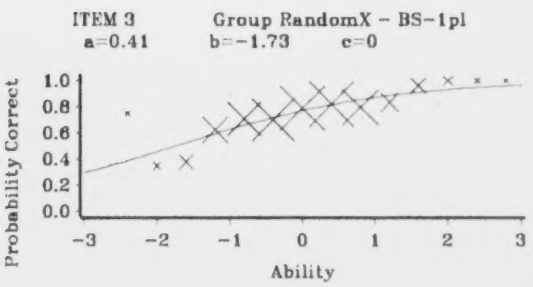


Figure 2b. Sample X - 2P

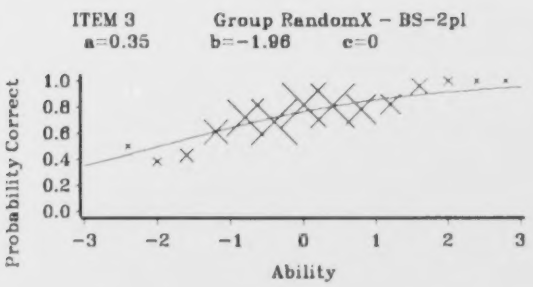


Figure 2c. Sample X - 3P

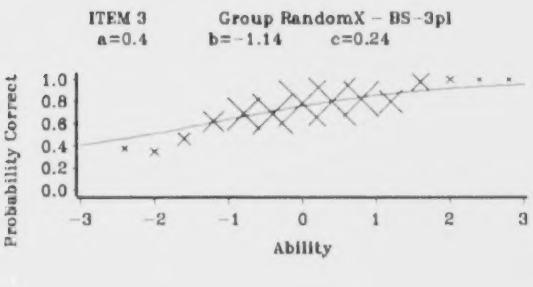


Figure 2d. Sample Female - 3P

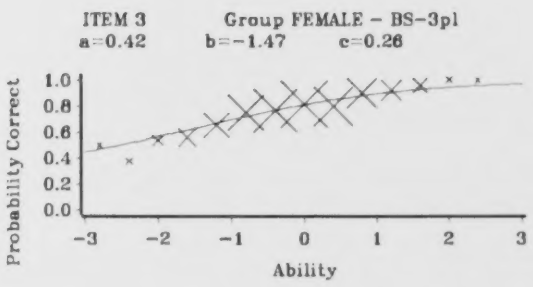


Figure 2e. Sample Asian- 3P

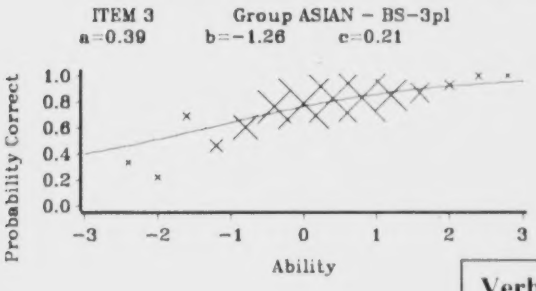
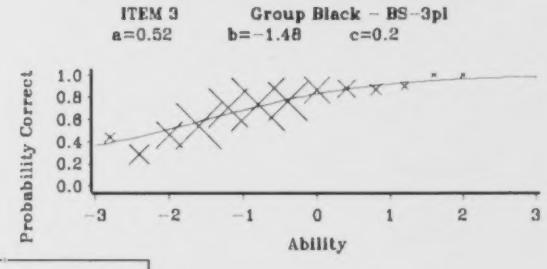


Figure 2f. Sample Black - 3P



Verbal Reasoning - item 3

Figure 2g. Sample X - 3P

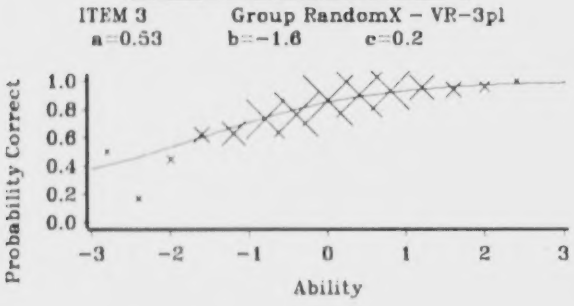
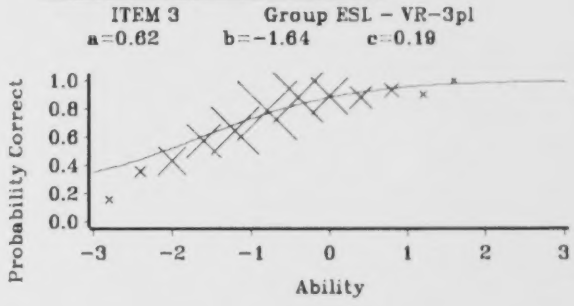


Figure 2h. Sample ESL - 3P



Figures 2g and 2h show 3P model/data fit for sample X and the ESL group respectively, for the Verbal Reasoning section.

Comparing across models, all three models studied (1P, 2P, and 3P) seem to fit the data (for representative sample X) fairly well. Model fit is best in the (-1, +1) proficiency interval, where most of the examinees lie (large crosses in this area). In the extreme intervals, especially in the lower range (-3, -1) the model does not always appear to fit as well. This apparent lack of fit is most likely due to very few examinees in this range (smaller sized crosses). All models underestimate in the higher ability range and overestimate in the lower intervals. The 2P model appears to fit somewhat better in the lower ability range.

The 3P Model fit for the female sample, in Figure 2d, is also good. This model fits the female response data slightly better in the lower proficiency range, than it does for the sample X data. This is most likely because, more people in this region probably leading to an improved fit. Figures 2e and 2f show the model/data fit for ethnic groups, Asian and black respectively. Again, the fit for the Asian subgroup is good in all proficiency intervals except in the lower range (-3, -1) , where there are few examinees. For the black subgroup, Figure 2f, the model/data fit is even adequate in the lower proficiency intervals where many examinees lie (large crosses). This item (BS- item 3) is also the most discriminating for the black group.

The 3P model/data fit of the ESL subgroup (Figure 2h) is compared to the fit for reference group X (Figure 2g) for item 3 in the VR test section. For this item, the model appears to fit the data well for all of the sample X examinees in the (-1, 3) ability range,

but fit is poorer in the lower ability intervals. Most ESL examinees lie in the (-2, +1) proficiency range for this item (VR - item 3) where there is good fit - fit is poor at both extreme ends.

Average Bias (Residual Error of Estimation). Average bias was evaluated separately for each item parameter for all models evaluated in this study. Tables 4a-4c show the results for the BS test section. Table 4a shows the bias in the slope parameter 'a' and its standard deviation for the Biological Science test section. Model 1P assumes fixed slope, hence 'a' is not included. Average bias for the random samples X/Y provides the base against which to compare the other bias estimates. There do not appear to be any systematic bias patterns across gender and racial/ethnic groups. Model 2P appears to have negatively biased 'a' estimates for the black group, indicating more discriminating items for this group. Model 3P 'a' estimates appear to have large positive bias for ESL, Male, and Asian subgroups, indicating less discriminating items for these groups as compared to group X. The white group has the largest negative bias in its 'a' estimates, suggesting that on average the items were more discriminating for this group. Comparing across models, Model 2P seems to provide the least biased 'a' parameter estimates (smaller standard deviations) for most gender and racial/ethnic groups.

Table 4b shows the average bias in 'b' parameter estimates, and their standard deviations, across models and samples. Comparing across models, the 1P model appears to provide the least biased estimates, with smallest variance, for the difficulty parameter 'b', than models 2P and 3P. Bias in the 1P estimates is almost negligible. Comparing across the gender and racial/ethnic groups for the 2P model , the 'b' estimate for the black and ESL groups (and the female to a lesser extent) is negatively biased as compared to

the base group X, suggesting that difficulty index was higher for these groups. The 'b' estimates for white subgroup was higher than the base, suggesting lower difficulty indices. For the 3P model, the sign of the bias was reversed as that compared to model 2P estimates, probably indicating an interaction with the additional parameter 'c' in this 3P model.

Table 4a.

Average bias in slope parameter 'a' - Biological Sciences (BS).

	Model	
Group	2P	3P
Random samples (XY)	-.004 (.06)	.008 (.15)
Asian	.002 (.07)	.037 (.14)
Black	-.026 (.12)	.017 (.18)
White	-.005 (.08)	-.029 (.15)
Female	.008 (.08)	-.004 (.13)
Male	-.001 (.06)	.042 (.16)
EPL	-.002 (.08)	.010 (.14)
ESL	.002 (.07)	.081 (.16)

Table 4b.

Average Bias in location parameter 'b' - Biological Sciences (BS).

	Model		
Group	1P	2P	3P
Random sample X/Y	-.002 (.15)	.010 (.23)	-.010 (.22)
Asian	.009 (.25)	.042 (.40)	-.001 (.40)
Black	-.007 (.32)	-.083 (.36)	.050 (.42)
White	.001 (.17)	.067 (.26)	-.067 (.23)
Female	-.000 (.16)	-.012 (.33)	-.024 (.44)
Male	.006 (.15)	.022 (.22)	.018 (.24)
EPL	.001 (.14)	.047 (.35)	.026 (.33)
ESL	.003 (.35)	-.031 (.35)	.088 (.39)

Table 4c.

Average Bias in intercept parameter 'c' - Biological Sciences (BS).

Group	Model 3P
Random Sample X/Y	.002 (.05)
Asian	.014 (.05)
Black	.015 (.06)
White	-.020 (.05)
Female	.002 (.05)
Male	.019 (.05)
EPL	.007 (.05)
ESL	.051 (.06)

Table 4c. shows the average bias in the "guessing" or intercept parameter 'c'. The 1P and 2P models do not include the c parameter. Compared to the base level of bias in random sample X/Y estimates, the positive bias in 'c' estimates is largest for the ESL group, suggesting lower 'c' values, on average, than those for sample X. Bias estimates are negative only for the white group, indicating larger average 'c' values for this group than the baseline.

Similar observations were made for the Physical Science test section, where, although the 'c' parameter estimates were slightly higher (negative bias) for all gender and racial/ethnic groups, while the ESL group had lower 'c' estimates, all bias estimates were within the random error range. For the Verbal Reasoning test section the average bias estimates were insignificant for all sampling groups, except the white group which had larger 'c' estimates (larger negative bias) than the base group. Average bias tables for the PS and VR test sections are presented in Appendix A.

Invariance of item parameter estimates. If the model fits the test data, parameter invariance usually follows, or if item parameter estimates are invariant - there is likely to be model/data fit. Item parameters, for each of the three models evaluated in this study, were estimated from the response data from all the samples. Parameters a, b, and c were estimated for the 3P model, parameters a and b for the 2P model and the item difficulty parameter 'b' alone for the 1P. Since parameter 'b' is the most critical of the three parameters, results are presented for the invariance of parameter 'b'. Invariance is established, if it can be shown that the scaled estimates obtained from calibrating the response data of various language, gender, and racial/ethnic groups are equivalent. Table 5 gives the correlations between the estimates for parameter 'b' obtained by calibrating the various groups.

Table 5.
Correlations for parameter 'b' for all 3 Models (1P, 2P, 3P) for tests (BS, PS, VR)

Sample	Biological Sciences (BS)			Physical Sciences (PS)			Verbal Reasoning (VR)		
	1P	2P	3P	1P	2P	3P	1P	2P	3P
X/Y	.994	.981	.984	.994	.978	.978	.993	.983	.975
Male/ Female	.988	.961	.934	.976	.945	.957	.978	.944	.953
White/ Asian	.971	.927	.951	.968	.940	.920	.980	.954	.945
White/ Black	.960	.934	.936	.960	.922	.951	.960	.945	.944
EPL/ESL	.958	.926	.935	.944	.924	.911	.966	.934	.949

Correlations between random samples X and Y serve as baseline for other comparisons. As expected, the X/Y correlations are the largest for all models and test

BEST COPY AVAILABLE

sections. However, regardless of the model or the test section, item parameter 'b' estimates from the various language, gender and racial/ethnic groups are all highly correlated (all correlations greater than .90). Correlations between Male/Female estimates are the highest, followed by White/Asian, White/Black and EPL/ESL. Comparing across models, all models appear to be supported by the IRT property of item parameter invariance.

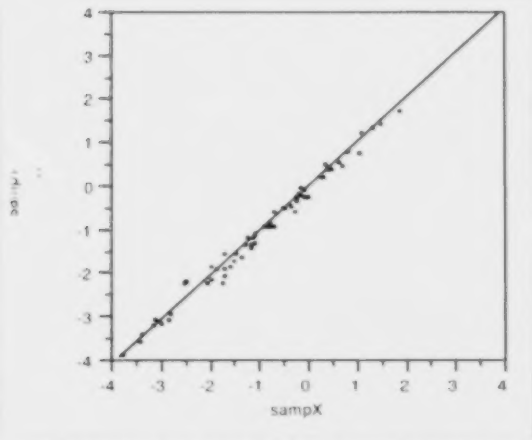


Fig. 3a. Plot of 'b'-1PL (BS) for samples X/Y (r=.994)

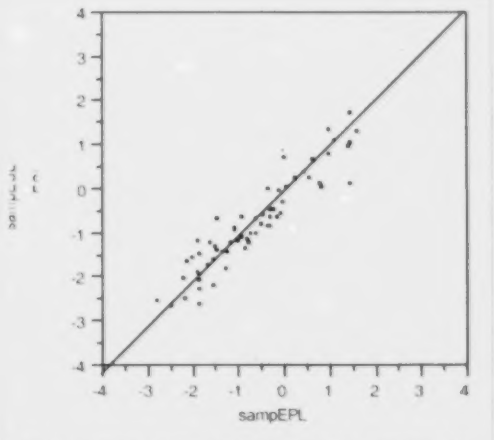


Fig. 3b. Plot of 'b'-2PL(BS) for EPL/ESL (r=.926)

Scatter plots illustrate the degree of model fit at various item difficulty levels. For the BS test section - the highest correlation is between the random samples X/Y for the 1P model and the lowest correlation is between the EPL/ESL group for the 2P model. Plots for these two samples/models for the BS test section are presented in Figures 3a and 3b. (The complete set of scatter plots for all samples, models and test sections are presented in the Appendix B). Plots between sample X and sample Y serve as baseline for other comparisons. The 45^0 or "identity" line is a reference line for judging the linear relationship between estimates. The scatter of the 'b' parameter estimates in the X/Y plot, Figure 3a., is tight around the 45^0 line, especially in the (-1, +1) difficulty range. The

BEST COPY AVAILABLE

scatter for the EPL/ESL parameter 'b' estimates is not as tight as for the sample X/Y plot. Even so, there appears to be a strong positive relationship between the EPL and ESL estimates.

Invariance of ability estimates. The stability and the distribution of the ability estimates for the various sampling groups assessed invariance of ability estimates with respect to gender, race/ethnic and language groups. This was done by comparing the distribution of ability estimates for each group, obtained using the calibrations from sample X, with those estimates of ability obtained from their own calibrations. This was done across models for model comparison. The mean ability estimates and their standard deviations for the 3 models (for the Biological Science test section) are presented in Table 6. (Mean estimates and standard deviations for the PS and VR test sections are presented in Appendix C). If the ability estimates, for all the language, gender and racial/ethnic groups from X-calibrations, are comparable to those obtained from their own respective calibrations, then invariance with respect to these groups will be established.

Using calibrations from sample X to obtain ability estimates for all the groups of interest, produced ability estimates that are consistently lower than those from same sample calibrations. This is observed across all models for all language, gender and racial/ethnic groups. The standard deviations for all estimates are between .68 and 1.076. The distribution of the ESL ability estimates appears to be the most heterogeneous. The largest difference between the two sets of ability estimates is observed for the black group across all 3 models. However, this difference is not large in standard deviation units. The ability estimates obtained using the 2P model are the closest, followed by 1P

and then 3P. In general, the two sets of estimates do appear to be equivalent for all models, supporting the invariance of ability parameters conclusion, with respect to the different groups.

Table 6.
Mean ability estimates for the various groups using their own calibrations and sample X calibrations - across the 3 IRT models- BS

	Model 1P		Model 2P		Model 3P	
Sample Groups	Grp-calib.	X-calib.	Grp-calib.	X-calib.	Grp-calib.	X-calib.
Asian	0.294	0.258	0.271	0.240	0.313	0.245
	(0.893)	(0.904)	(0.902)	(0.902)	(0.849)	(0.884)
Black	-0.990	-0.855	-0.939	-0.839	-0.988	-0.860
	(0.746)	(0.813)	(0.678)	(0.782)	(0.884)	(0.85)
White	0.296	0.261	0.271	0.251	0.319	0.256
	(0.845)	(0.870)	(0.860)	(0.873)	(0.787)	(0.841)
Female	-0.149	-0.123	-0.154	-0.121	-0.120	-0.121
	(0.964)	(0.966)	(0.952)	(0.955)	(0.969)	(0.958)
Male	0.271	0.235	0.255	0.224	0.285	0.217
	(0.969)	(0.954)	(0.971)	(0.950)	(0.932)	(0.929)
EPL	0.119	0.108	0.106	0.103	0.139	0.105
	(0.980)	(0.968)	(0.978)	(0.958)	(0.949)	(0.944)
ESL	-0.083	-0.067	-0.081	-0.079	-0.059	-0.090
	(1.075)	(1.042)	(1.052)	(1.023)	(1.076)	(1.032)

Invariance of ability parameter estimates for tests that differed in difficulty was assessed next. This was done by comparing ability estimates obtained from artificially created hard and easy tests. Hard tests for the BS and PS test sections consisted of 30 hardest items from the total test (25 for VR). Easy tests consisted of the 30 easiest items from the total test (25 for VR). Two tests of equivalent difficulty, one consisting of the odd half and the other consisting of the even half of the total test, served as reference or baseline tests against which to compare hard/easy test estimates. The ability estimates for each sample/test were obtained using calibrations from sample X; these estimates are

BEST COPY AVAILABLE

presented in Table 7. The ability estimates from the hard/easy tests were compared to the estimates from the total test. Again this was done across all the language, gender and racial/ethnic groups and models. In general, the estimates obtained from hard/easy tests are comparable to those from the total test.

Table 7.
Mean ability estimates and standard deviations for all subgroups, from hard, easy and total test - BS.

	Model 1P			Model 2P			Model 3P		
Sample	Hard	Easy	Total	Hard	Easy	Total	Hard	Easy	Total
X	0.012	0.022	0.009	0.004	-0.002	0.003	0.012	0.008	0.008
	(0.802)	(0.932)	(0.940)	(0.853)	(0.863)	(0.932)	(0.831)	(0.876)	(.934)
Asian	0.227	0.240	0.258	0.231	0.191	0.240	0.230	0.204	0.245
	(0.801)	(0.872)	(0.904)	(0.849)	(0.818)	(0.902)	(0.831)	(0.825)	(0.884)
Black	-0.618	-0.825	-0.855	-0.672	-0.775	-0.839	-0.630	-0.782	-0.860
	(0.636)	(0.931)	(0.813)	(0.668)	(0.839)	(0.782)	(0.620)	(0.860)	(0.85)
White	0.186	0.280	0.261	0.191	0.235	0.251	0.194	0.248	0.256
	(0.781)	(0.838)	(0.870)	(0.836)	(0.788)	(0.873)	(0.808)	(0.793)	(0.841)
Female	-0.114	-0.077	-0.123	-0.127	-0.093	-0.121	-0.113	-0.085	-0.121
	(0.805)	(0.992)	(0.966)	(0.859)	(0.913)	(0.955)	(0.826)	(0.926)	(0.958)
Male	0.200	0.222	0.235	0.203	0.185	0.224	0.205	0.196	0.217
	(0.823)	(0.924)	(0.954)	(0.874)	(0.861)	(0.950)	(0.839)	(0.870)	(0.929)
EPL	0.071	0.122	0.108	0.071	0.091	0.103	0.078	0.101	0.105
	(0.841)	(0.943)	(0.968)	(0.897)	(0.872)	(0.958)	(0.865)	(0.883)	(0.944)
ESL	0.023	-0.112	-0.067	0.008	-0.122	-0.079	0.013	-0.116	-0.090
	(0.839)	(1.044)	(1.042)	(0.889)	(0.959)	(1.023)	(0.854)	(0.975)	(1.032)

However, estimates from the easy tests are closer in magnitude to the total test estimates. There does not appear to be any systematic pattern in the distribution of these estimates across models. For the ESL subgroup, the estimates from the hard tests are positive (for all models) as compared to the easy and total test estimates, though the difference is not large in standard deviation units. (Mean ability estimates for the PS and VR test sections are in Appendix D).

BEST COPY AVAILABLE

Table 8a, gives the attenuated correlations between ability estimates obtained using odd/even tests and correlations between ability estimates obtained using hard/easy tests. Correlations for the Biological Sciences test section only are reported in this table.

Table 8a. Attenuated correlations between ability estimates - BS

	Odd-Even			Hard-Easy		
Sample	3 PL	2 PL	1 PL	3 PL	2 PL	1 PL
Random X	.80	.79	.76	.72	.72	.70
Asian	.76	.76	.74	.71	.71	.70
Black	.74	.74	.70	.62	.61	.60
White	.75	.75	.73	.69	.69	.68
Female	.79	.79	.77	.73	.72	.71
Male	.79	.78	.77	.73	.73	.72
EPL	.79	.78	.77	.74	.74	.72
ESL	.82	.82	.80	.77	.77	.76

The correlations in Table 8a are lower than those obtained in other analyses reported in this study, due to the shorter test lengths (30 items for BS and PS, and 25 item for the VR test section) used to estimate ability. The reliability for the whole test was 0.85 (which serves as an upper bound for any of the attenuated test correlations), while that for the shorter test was 0.75. To estimate the strength of the linear relationship that might be observed, if we did not have the problem of unreliability in the shortened tests, the correlations in Table 8a have been disattenuated (corrected) using the Spearman-Brown Prophecy.

These corrected correlations are presented in Table 8b. Correlations between ability estimates from odd/even tests seem to be very compatible with those obtained from the hard/easy tests suggesting invariance of ability parameter estimates with respect to test difficulty. (Correlation tables for the other two sections can be found in the Appendix E).

BEST COPY AVAILABLE

Table 8b.

Disattenuated correlations between ability estimates - BS

	Odd-Even			Hard-Easy		
Sample	3 P	2 P	1 P	3 P	2 P	1 P
Random X	0.89	0.89	0.87	0.84	0.84	0.83
Asian	0.87	0.87	0.86	0.84	0.84	0.83
Black	0.86	0.86	0.83	0.77	0.77	0.76
White	0.86	0.86	0.85	0.82	0.82	0.82
Female	0.89	0.89	0.88	0.85	0.84	0.84
Male	0.89	0.88	0.88	0.85	0.85	0.84
EPL	0.89	0.88	0.88	0.86	0.86	0.84
ESL	0.91	0.91	0.89	0.88	0.88	0.87

Although, ability estimate correlations, for the racial/ethnic groups, between hard/easy tests are consistently lower than those between baseline odd/even test, the magnitude of the difference in correlations is small and not of any practical significance. The correlations between the 2P estimates are very similar to the 3P estimate correlations. Those for the 1P model are slightly lower, although again the magnitude of the difference is very small and not of practical significance. This pattern for the odd/even and the hard/easy test correlations is observed across samples and subtests.

This equivalence between ability estimate correlations from odd/even tests and hard/easy tests does support the conclusion of invariance of ability parameter estimates with respect to test difficulty.

Model Comparison. Model comparisons were incorporated in the analyses of model fit and parameter invariance. Some additional techniques for model comparisons are presented in this section. One way to address the utility of the 1P model as compared to the more complex 2P and 3P models is to look at a plot of the slope or discrimination

BEST COPY AVAILABLE

parameter 'a' versus the item difficulty parameter 'b'. An example of this a/b plot (for the BS test section) is presented in Figure 4. (Plots for the MS and VR sections are in

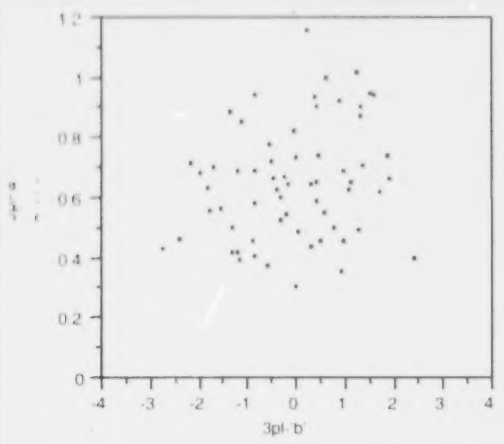


Fig.4. Plot of parameters a/b for 3P (sample X - BS)

Appendix F). The range of values for the discrimination parameter 'a' estimates should be small to support the choice of a 1P model, since this model assumes equal discrimination. The a/b plot for the 3P model, Figure 4, shows that the values of slope parameter 'a' estimates vary for all values of 'b' and that this variation is random. In the difficulty range (-2, +2) where most of the items (and examinees) lie, the parameter estimates for 'a' vary from .3 to 1.2 for the BS test section, suggesting that all items are not equally discriminating. This may signify a need for the inclusion of the slope parameter in our model to improve model fit at the item level.

Comparing the ability estimates for the random sample X, obtained using calibrations from the various language, gender, and racial/ethnic groups, across models assessed stability of the parameter estimates across models and groups. Correlations between ability estimates for sample X obtained using sample X calibrations, and the corresponding ability estimates for sample X obtained using calibrated item statistics

based upon the other groups are reported in Table 9. These are presented across the three models for comparison purposes. (Tables of correlations for the PS and VR test sections are presented in the Appendix H). The mean ability estimates for sample X, obtained from the different calibrations are in Table 10. (Corresponding estimates for PS and VR test sections are in Appendix H).

Table 9.
Correlations between theta estimates for sample X obtained using sample X calibrations and estimates obtained using calibrations from other subgroups - (BS).

Calibration Group	Model 1P	2P	3P
Asian	.9999	.9984	.9973
Black	.9999	.9952	.9958
White	.9999	.9979	.9976
Female	.9999	.9984	.9977
Male	.9999	.9987	.9973
Epl	.9999	.9985	.9978
Esl	.9999	.9981	.9961

All correlations in Table 9, across models and calibration samples are extremely high, greater than 0.99. All 3 models under investigation are primarily based on the proportion correct score, this score remains constant for sample X, hence, irrespective of the calibrations used, the ability estimates would all be linearly related. This is most striking for the 1P model, which depends solely on the item difficulty (proportion correct score). The other models 2P and 3P, which incorporate the discrimination parameter 'a' and intercept 'c', also offer estimates that are closely related. This is also supported by the distribution of ability estimates in Table 10. The ability estimates do appear to be stable regardless of the calibrations used to obtain them. The estimates obtained using the black group calibrations differ the most from the sample X estimates, although even this difference is slight (in SD units) and not of practical significance.

Table 10.
Mean ability estimates (standard deviations) for sample X obtained using calibrations from various groups - BS

Calibration Group	Mean ability estimates for Sample X (sd)		
	1P	2P	3P
Sample X	0.009 (0.940)	0.003 (0.932)	0.008 (0.934)
Asian	0.009 (0.941)	-0.001 (0.920)	0.015 (0.914)
Black	0.009 (0.937)	-0.033 (0.903)	-0.017 (0.922)
White	0.009 (0.943)	0.003 (0.916)	0.011 (0.915)
Female	0.009 (0.941)	-0.007 (0.920)	-0.003 (0.924)
Male	0.009 (0.941)	0.007 (0.923)	0.020 (0.920)
EPL	0.009 (0.942)	-0.003 (0.921)	0.004 (0.922)
ESL	0.009 (0.938)	-0.000 (0.923)	0.018 (0.931)

Plots of the empirical standard errors, with 95% error bars, versus proficiency (ability) interval midpoints for the Biological Science subtest are presented in Figures 5a and 5b. These plots help us compare between models relative to the standard errors of ability estimation at different ability levels. Plots for the white and black subgroups are presented here; SE plots for all other samples for the BS test section are in the Appendix I.

The distribution of standard errors appears to be similar for all the models, 1P, 2P, and 3P in this study. Standard errors are large for intervals with fewer examinees. The 1P standard errors appear to be slightly larger in the lower proficiency intervals in most groups (except in the black group- where the 3P errors are largest) probably due to the small number of examinees in that interval. The 2P standard errors are the lowest in all groups except in some of the highest ability intervals, where they are only slightly higher or equivalent to the other models.

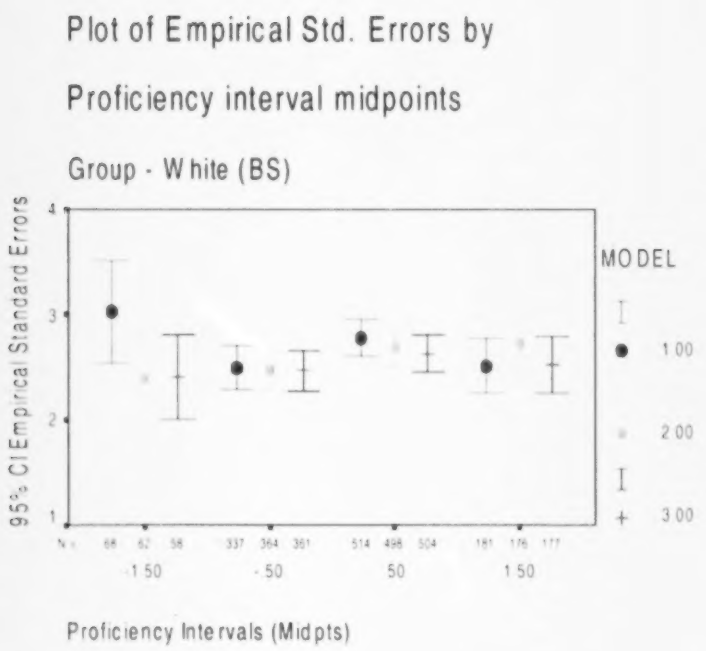


Figure 5a. Empirical Standard Errors for the White subgroup - BS

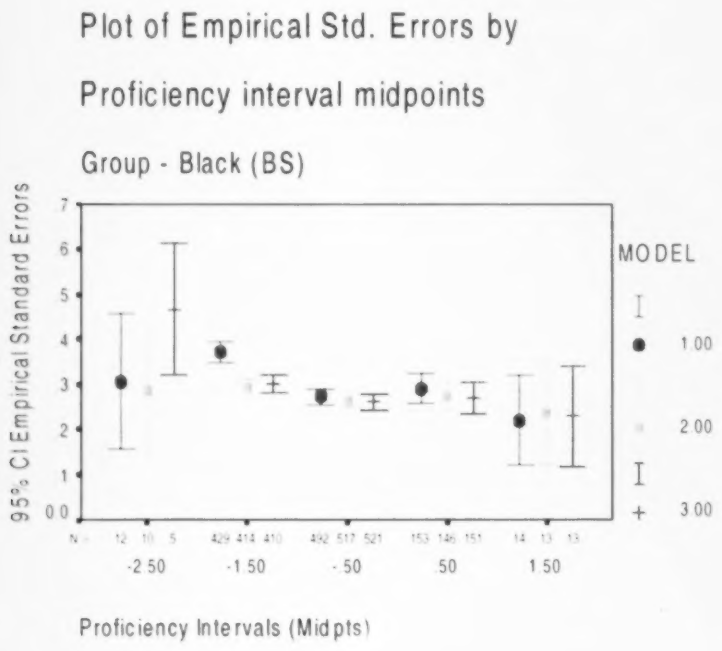


Figure 5b. Empirical Standard Errors for the Black subgroup - BS

BEST COPY AVAILABLE

Discussion and Conclusion

The purpose of this study was to investigate the viability of the property of parameter invariance for the 1P, 2P and the 3P item response models. Invariance of item parameters across different gender, ethnic and language groups, and the invariance of ability parameters with respect to test difficulty was assessed. It also sought to test the stability of ability estimates obtained for random sample X using calibrations from different groups and to identify the most efficient IRT model suitable for the MCAT data.

Limitations. A number of factors limit the results of this investigation of IRT parameter invariance for the MCAT data. First, the 3 test sections were treated as independent tests and 9 different samples were drawn from each test section, rather than one set of 9 samples from the total test taker population. This is likely to be a source of additional sampling error and it limits model/parameter comparisons across test sections.

Second, the overlap in the baseline X and Y samples, although slight, gives a lower estimate of the normal sampling error expected in this study, against which other effects are compared. Third, this study uses only samples of size 1,100 - since this is adequate (lower bound) for fitting the 3P model. In many computer based testing situations, the pretest/online calibrations are based on much smaller sample sizes. Hence, this sample size of 1,100 does not provide information on the relative impact for smaller sample sizes.

Finally, using BILOG 3.11 for calibration resulted in only Bayesian parameter estimates being considered.

Future Research As MCAT considers this important leap from Paper & Pencil (P&P) testing to Computer Based Testing (CBT) here are some directions for future research. Since this study does not offer any conclusive evidence on model selection, this investigation should be extended to include item/test information functions and residual analyses to further aid in the selection of an appropriate model.

As discussed earlier, in practical situations, pretesting for item calibration is usually done with much smaller sample sizes than that considered here (1,100), hence, a replication of this study with smaller sample sizes will be useful in assessing the stability of estimates based on smaller samples.

After the initial pretest calibrations, new item additions to the item pool are calibrated on-line. Small initial errors in parameter estimates can accumulate over time to systematically distort the score scale. Research on item/scale score drift will lead to more reliable score estimates. Estimation procedures that will minimize scale score distortions and lead to stable score estimates over time should also be investigated.

Once the groundwork for CBT has been done, choice of CAT models, and then based upon the dimensionality studies - constraints on item selection algorithms need to be established. Based upon dimensionality studies, the need for multidimensional IRT models should also be explored

Conclusion. Since the assumption of unidimensionality is critical for any IRT model/data fit, this assumption was first tested. Although this assumption cannot be completely met by any test data, it suffices to show that there exists only one "dominant" factor that influences the test performance. The present study reported the presence of two or more underlying dimensions or factors affecting test performance. Future

research regarding the sources of this multidimensionality needs to be done. A study by Childs and Oppler (1999), tested the two disciplines within the BS (Biology and Organic Chemistry) and PS (Physics and General Chemistry) test sections as possible sources to violations of the unidimensionality assumption. In their study, regarding the practical implications of subtest dimensionality of the MCAT, Childs and Oppler (1999) conclude "...that the discipline-based multidimensionality of the science subtests may be somewhat immaterial in the calibration of the MCAT item bank". Further research needs to be done regarding this assumption. If considering the implementation of computer adaptive testing, caution must be exercised since this assumption of unidimensionality is critical in creating item pools for the CAT.

Three IRT models were then fit to the dichotomous response data from every sample. All three models tested, show adequate fit for the MCAT data. The 1P model-item estimates had the smallest estimation error (average bias) as compared to the 2P and 3P, although the difference was very small in magnitude.

There is evidence to support the conclusion that item and ability parameters are stable/invariant with respect to gender, racial/ethnic and language groups for all models. Ability estimates also appear to be invariant with respect to test difficulty for all models.

Further, the ability estimates were also found to be stable across calibration samples for all models. The plots of parameter 'a' against parameter 'b', were prepared to assess the adequacy of the 1P model, which assumes constant 'a's. These plots revealed large variance in slopes or discrimination at all levels of difficulty 'b'. However this variation was more uniform than systematic suggesting a need for caution in a

decision to include/exclude this parameter. The plots of the empirical standard errors do support the adequacy of the simpler 1P or 2P models for the MCAT data.

Although this study supports an IRT model fit for the MCAT data, and establishes parameter invariance, the decision regarding the selection of an appropriate model to fit the MCAT data is still inconclusive.

References

Childs, R. A., & Oppler, S. H. (1999). Practical implications of subtest dimensionality for item response theory calibration of the Medical College Admission Test. Research Report; American Institutes for Research. Washington, DC.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principals and applications.* Boston: Kluwer.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26, 195-211.*

Kingston, N.M., & Dorans, N. J. (1985). The analysis of item-ability regressions: an exploratory IRT model fit tool. *Applied Psychological Measurement, 9, 281-288.*

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, Mass: Addison-Wesley.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27, 82-89.*

Meara, K., & Sireci, S. G. (1998). Appraising the dimensionality of the MCAT. *MCAT/GSRP Final Report.*

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics,* 4, 207-230.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,* 7, pp. 201-210.

Appendix A

Table A1.

Average Bias in slope parameter 'a' for all samples for all three models - Physical Sciences (PS).

	Model	
Group	2P	3P
Random Sample X/Y	-.006 (.08)	.011 (.13)
Asian	-.002 (.08)	-.039 (.14)
Black	-.019 (.14)	-.031 (.15)
White	-.003 (.07)	-.048 (.14)
Female	-.009 (.09)	-.051 (.13)
Male	.001 (.06)	-.005 (.10)
EPL	-.011 (.09)	-.042 (.11)
ESL	.016 (.07)	.021 (.13)

Table A2.

Average Bias in location parameter 'b' for all samples for all three models - Physical Sciences (PS).

		Model	
Group	1P	2P	3P
Random Sample X/ Y	.005 (.13)	-.030 (.28)	.037 (.24)
Asian	-.004 (.25)	.023 (.34)	-.085 (.39)
Black	.010 (.34)	-.217 (.43)	-.035 (.27)
White	.008 (.15)	.111 (.35)	.012 (.28)
Female	-.002 (.19)	-.119 (.40)	-.083 (.25)
Male	.009 (.17)	.020 (.26)	-.043 (.26)
EPL	.014 (.18)	-.027 (.24)	-.059 (.23)
ESL	.001 (.38)	-.043 (.45)	-.038 (.45)

Table A3.

Average Bias in intercept parameter 'c' for all samples for all three models - Physical Sciences (PS).

	Model
Group	3P
Random Sample X/Y	.021 (.03)
Asian	-.020 (.04)
Black	-.006 (.04)
White	-.020 (.04)
Female	-.019 (.04)
Male	-.007 (.04)
EPL	-.023 (.04)
ESL	.001 (.05)

Table A4.
Average Bias in slope parameter 'a' for all samples for all three models - Verbal Reasoning (VR).

	Model	
Group	2P	3P
Random Sample X/Y	.000 (.08)	-.037 (.13)
Asian	.002 (.07)	-.013 (.10)
Black	-.008 (.11)	-.072 (.22)
White	-.001 (.09)	-.105 (.14)
Female	-.002 (.07)	.003 (.10)
Male	-.001 (.09)	-.018 (.13)
EPL	-.001 (.08)	.009 (.12)
ESL	-.002 (.12)	.007 (.14)

Table A5.
Average Bias in location parameter 'b' for all samples for all three models - Verbal Reasoning (VR).

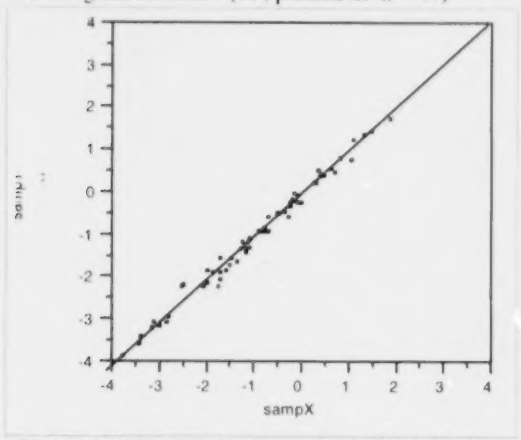
		Model	
Group	1P	2P	3P
Random Sample X/Y	-.000 (.13)	.010 (.34)	-.087 (.33)
Asian	-.001 (.15)	-.004 (.25)	-.016 (.29)
Black	-.001 (.24)	-.093 (.26)	-.085 (.32)
White	-.006 (.15)	-.020 (.24)	-.079 (.25)
Female	.001 (.16)	.007 (.26)	.016 (.25)
Male	-.007 (.13)	-.008 (.22)	-.054 (.23)
EPL	-.003 (.10)	.000 (.26)	.009 (.22)
ESL	-.004 (.28)	-.070 (.31)	.009 (.34)

Table A6.
Average Bias in intercept parameter 'c' for all samples for all three models -Verbal Reasoning (VR).

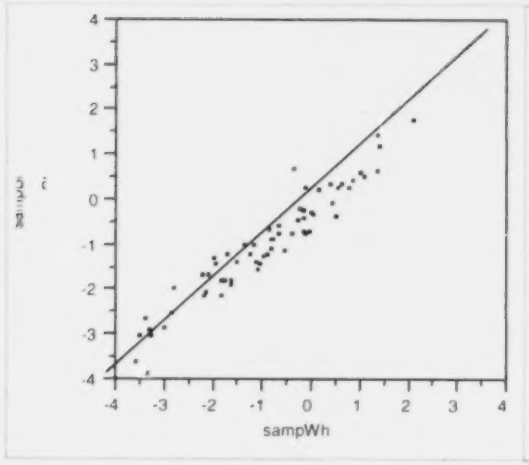
	Model
Group	3P
Random Sample X/ Y	-.037 (.05)
Asian	-.009 (.04)
Black	-.040 (.07)
White	-.089 (.05)
Female	.006 (.04)
Male	-.016 (.04)
EPL	.007 (.04)
ESL	.004 (.04)

Appendix B

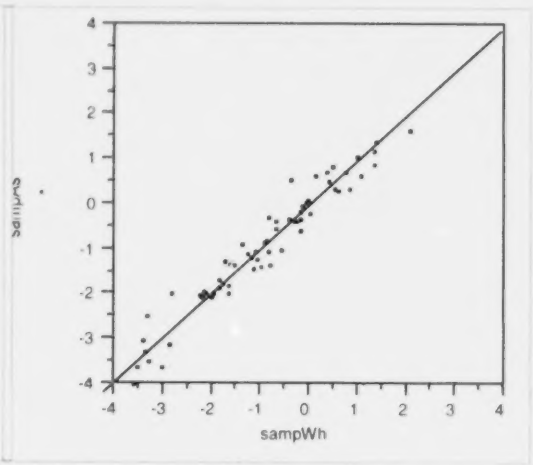
Biological Sciences (1P, parameter 'a' = .4)



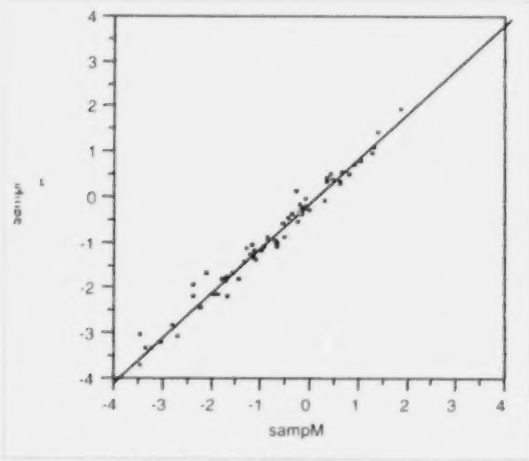
Plot of 'b'-1P (BS) for Samples X/Y (r=.994)



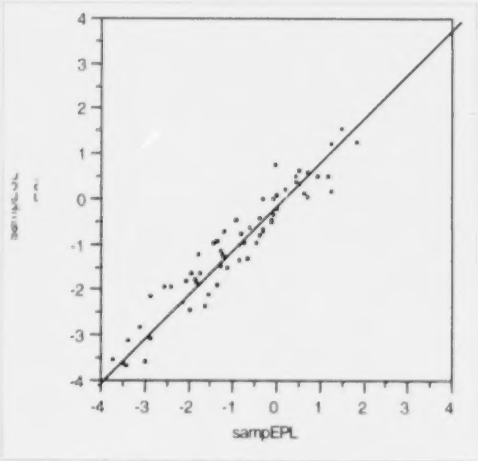
Plot of 'b'-1P for Wh/As (r=.971)



Plot of 'b'-1P for Wh/Bl (r=.960)

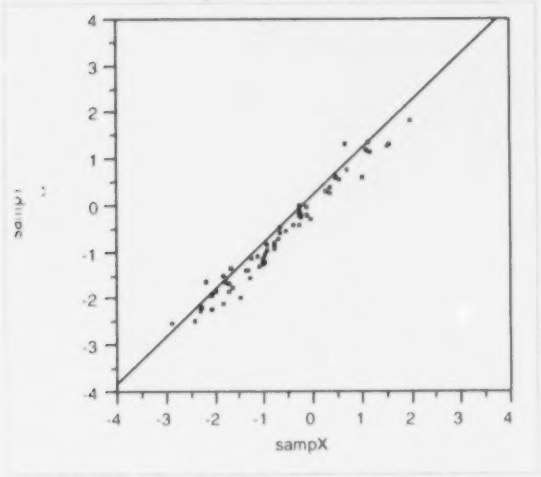


Plot of 'b'-1P for M/F (r=.988)

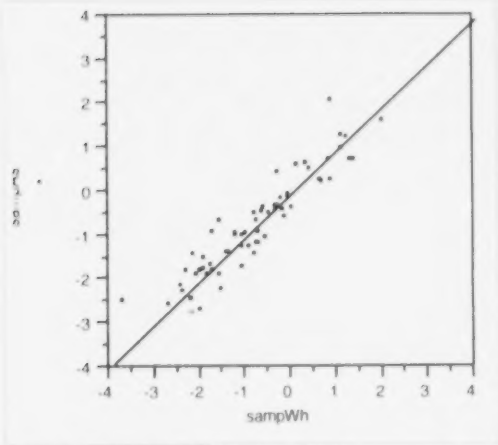


Plot of 'b'-1P for EP/ESL (r=.958)

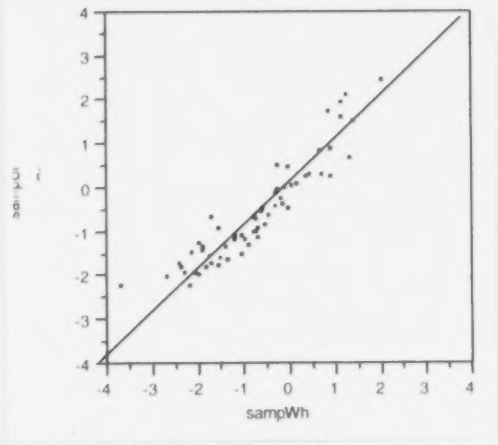
Biological Sciences (2P)



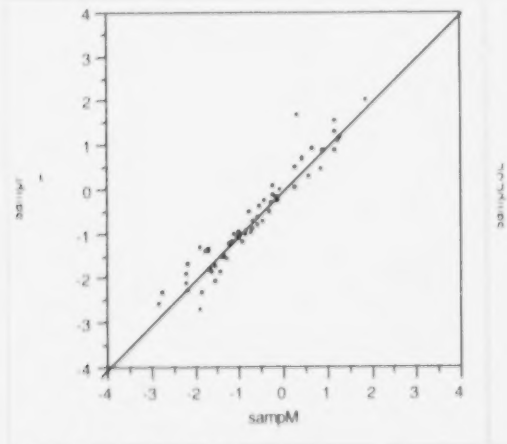
Plot of 'b'-2P for Samples X/Y (r=.981)



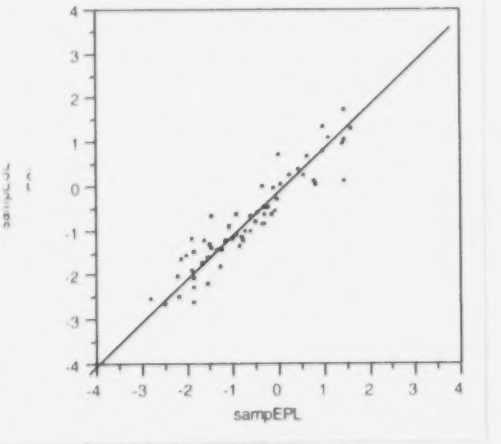
Plot of 'b'-2P for Wh/As (r=.927)



Plot of 'b'-2P for Wh/Bl (r=.934)

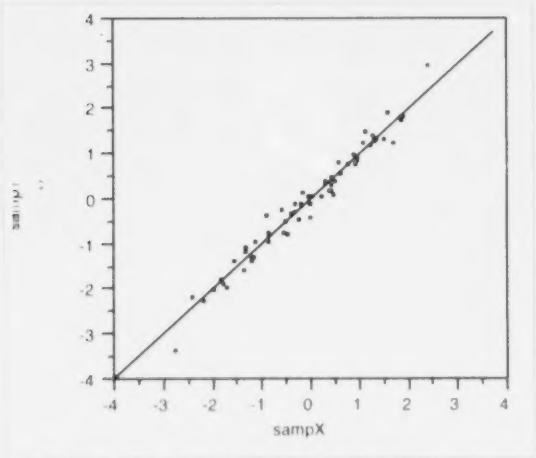


. Plot of 'b'-2P for M/F (r=.961)

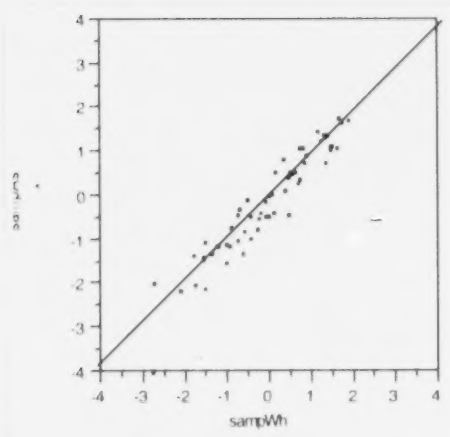


Plot of 'b'-2P for EP/ESL (r=.926)

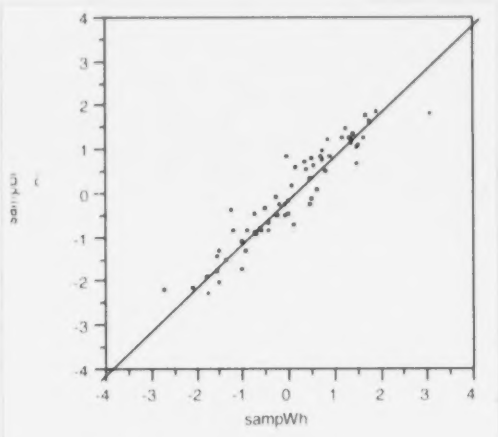
Biological Sciences (3P)



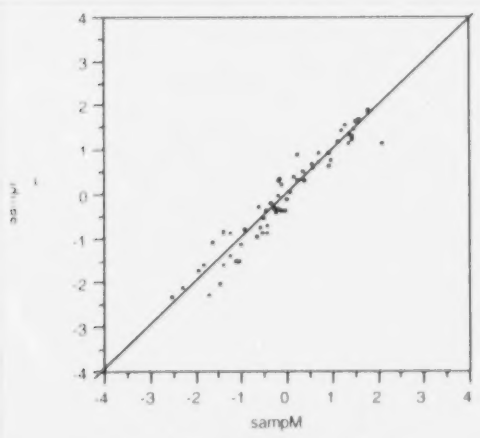
Plot of 'b'-3P for Samples X/Y (r=.984)



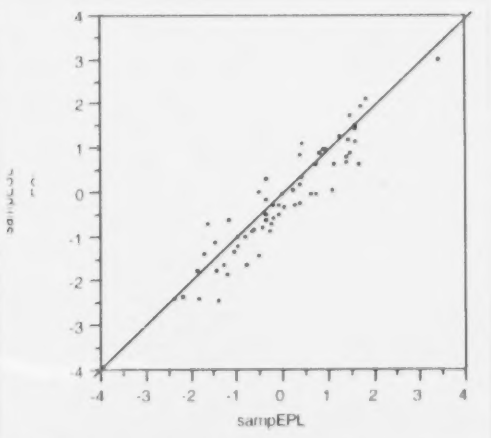
Plot of 'b'-3P for Wh/As (r=.951)



Plot of 'b'-3P for Wh/Bl (r=.936)

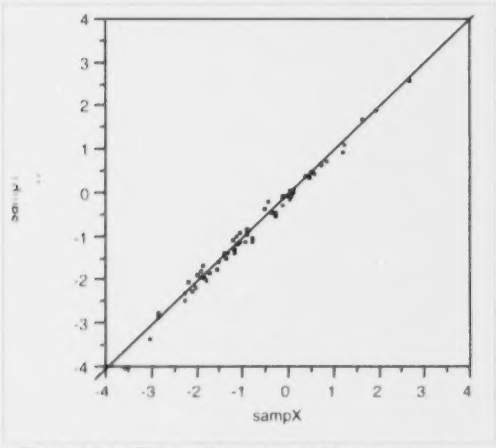


Plot of 'b'-3P for M/F (r=.964)

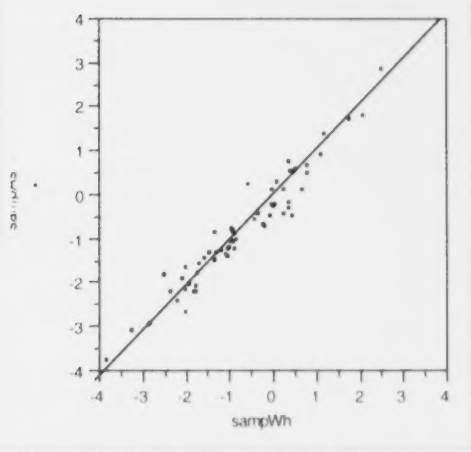


Plot of 'b'-3P for EP/ESL (r=.935)

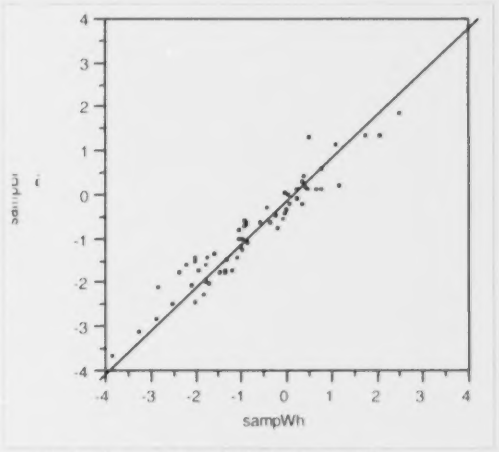
Physical Sciences (1P, parameter 'a' = .45)



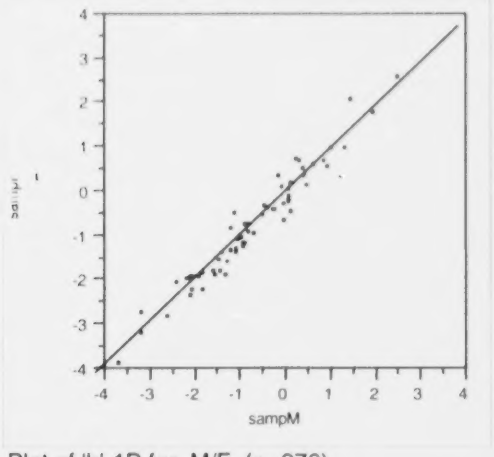
Plot of 'b'-1P for Samples X/Y (r=.995)



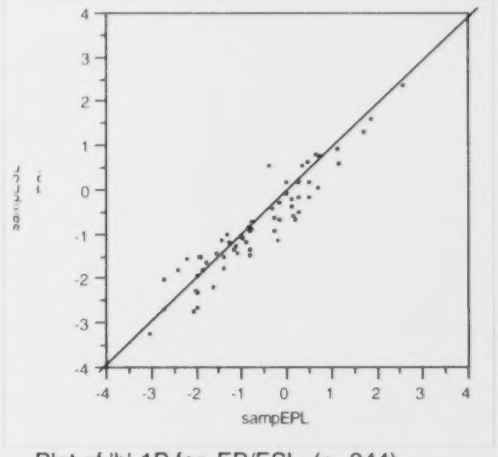
Plot of 'b'-1P for Wh/As (r=.968)



Plot of 'b'-1P for Wh/Bl (r=.960)



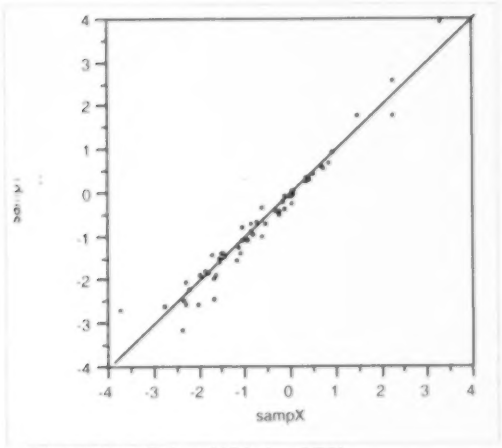
Plot of 'b'-1P for M/F (r=.976)



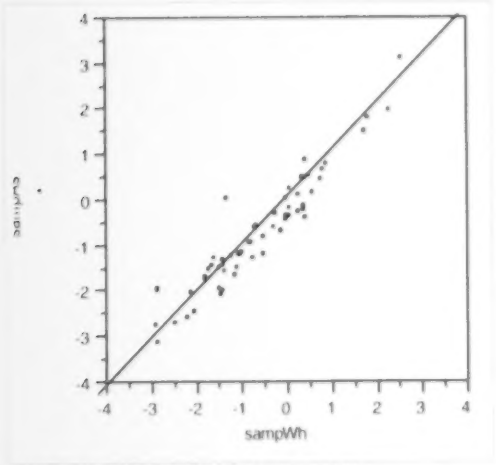
Plot of 'b'-1P for EP/ESL (r=.944)

BEST COPY AVAILABLE

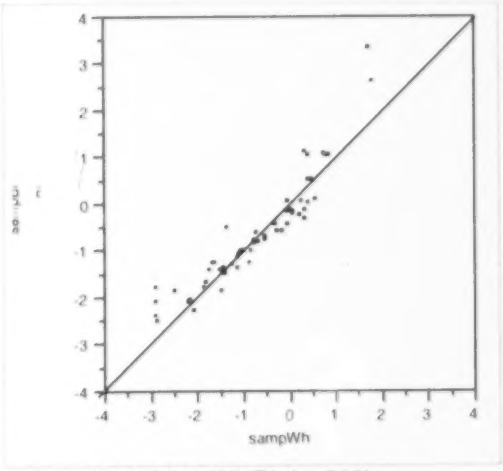
Physical Sciences (2P)



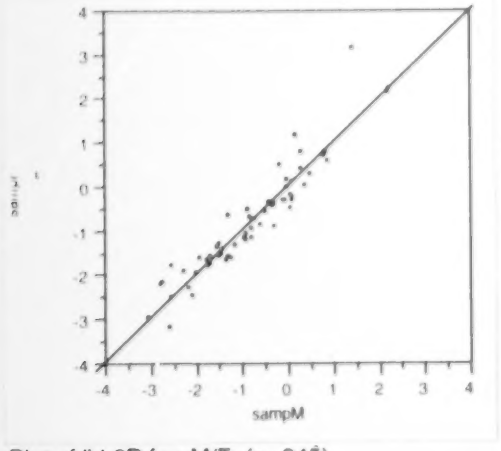
Plot of 'b'-2P for X/Y (r=.978)



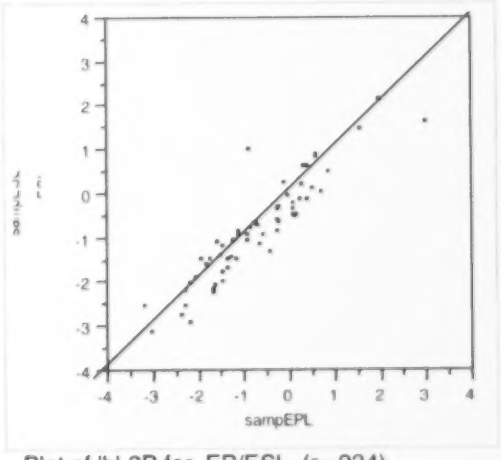
Plot of 'b'-2P for Wh/As (r=.940)



Plot of 'b'-2P for Wh/Bl (r=.922)

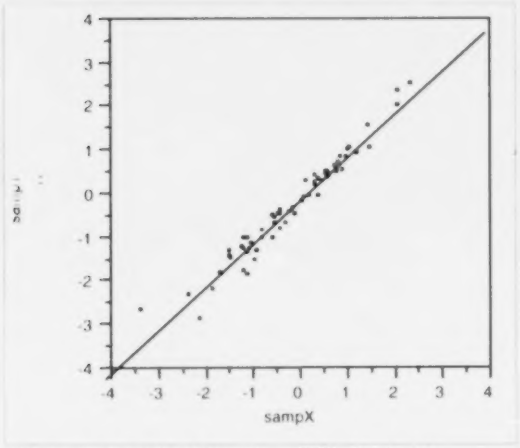


Plot of 'b'-2P for M/F (r=.945)

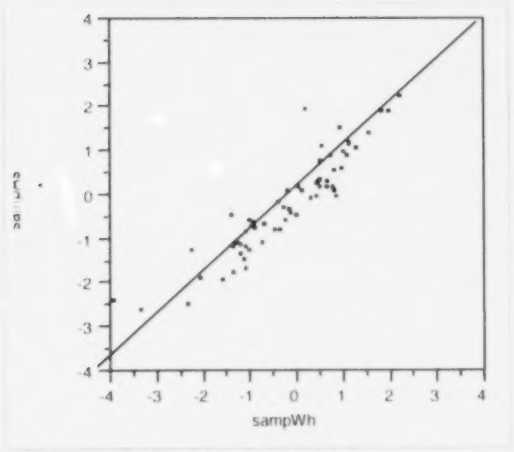


Plot of 'b'-2P for EP/ESL (r=.924)

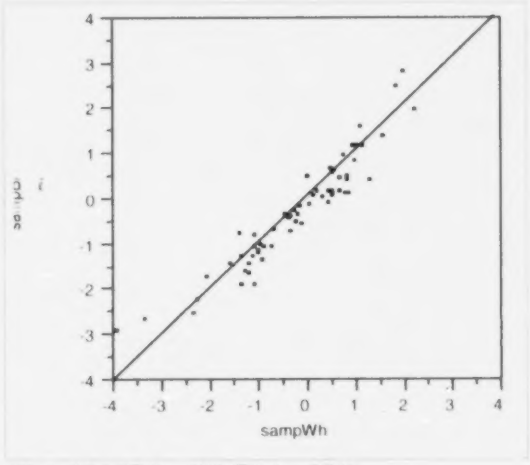
Physical Sciences (3P)



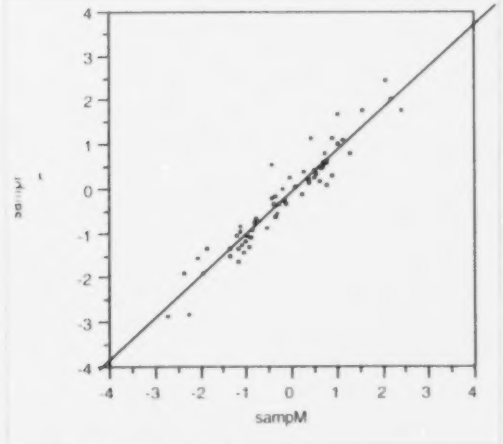
Plot of 'b'-3P for Samples X/Y (r=.978)



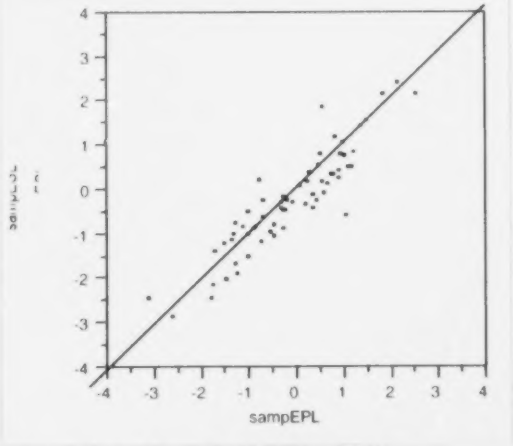
Plot of 'b'-3P for Wh/As (r=.920)



Plot of 'b'-3P for Wh/Bl (r=.951)

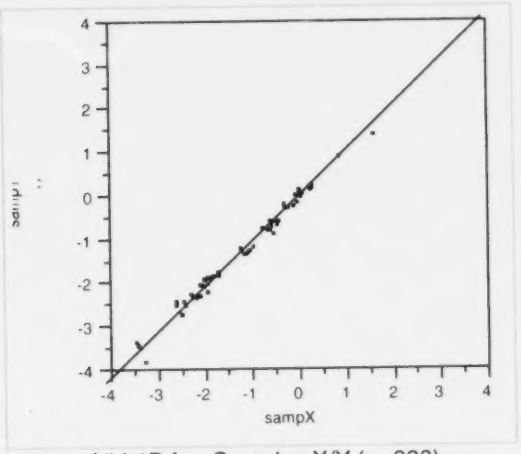


Plot of 'b'-3P for M/F (r=.957)

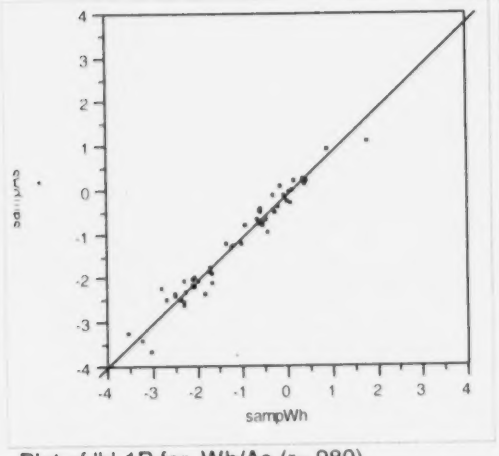


Plot of 'b'-3P for EP/ESL (r=.911)

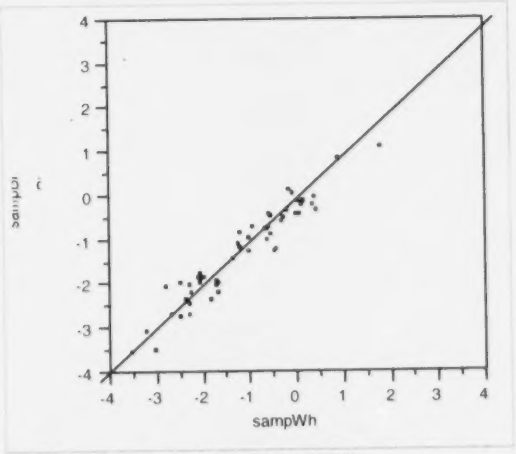
Verbal Reasoning (1P, parameter 'a' = .47)



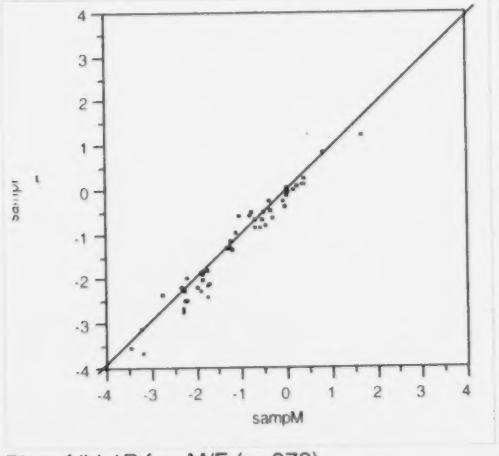
Plot of 'b'-1P for Samples X/Y (r=.993)



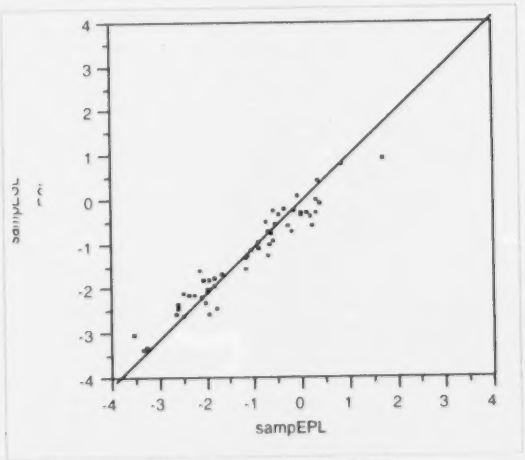
Plot of 'b'-1P for Wh/As (r=.980)



Plot of 'b'-1P for Wh/Bl (r=.960)

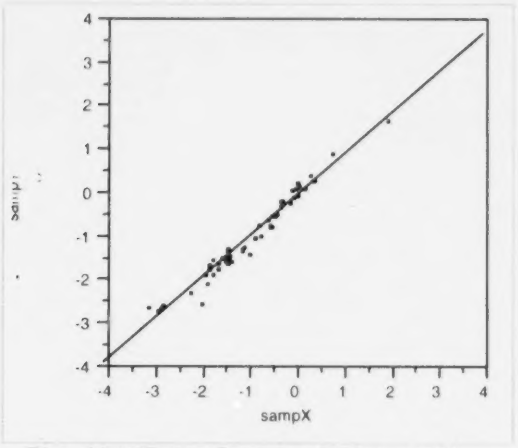


Plot of 'b'-1P for M/F (r=.978)

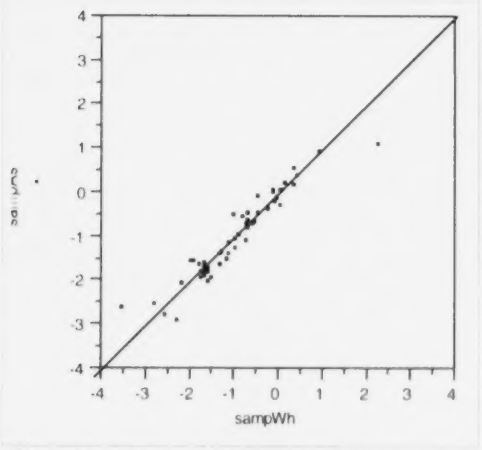


Plot of 'b'-1P for EP/ESL (r=.966)

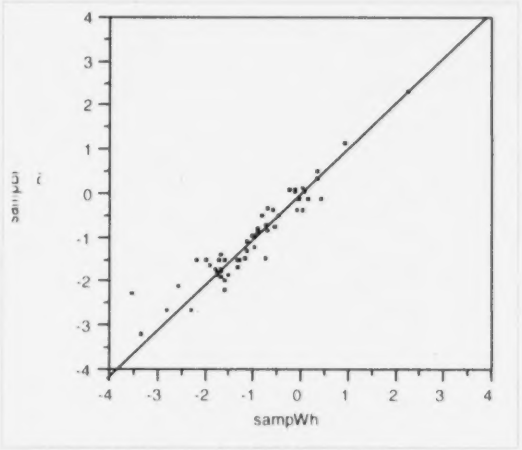
Verbal Reasoning (2P)



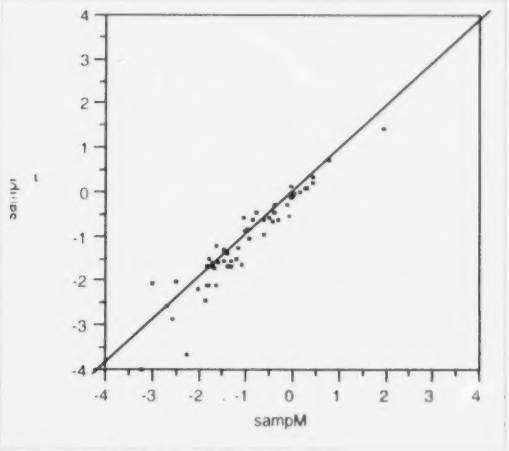
Plot of 'b'-2P for Samples X/Y (r=.983)



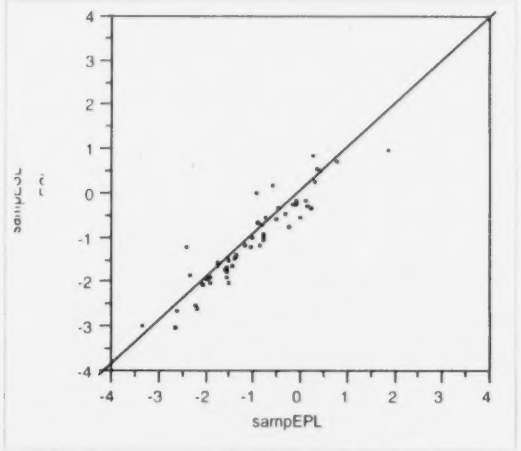
Plot of 'b'-2P for Wh/As (r=.954)



Plot of 'b'-2P for Wh/Bl (r=.949)

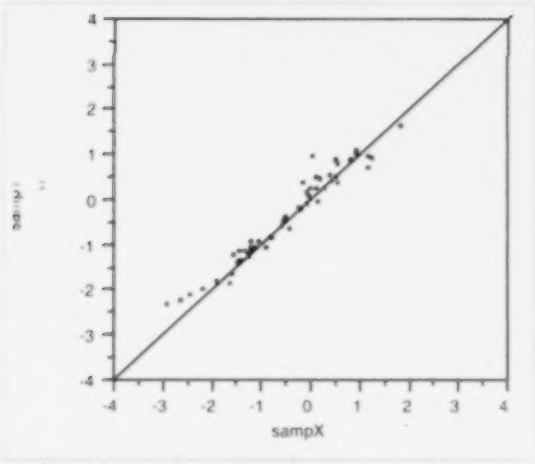


Plot of 'b'-2P for M/F (r=.944)

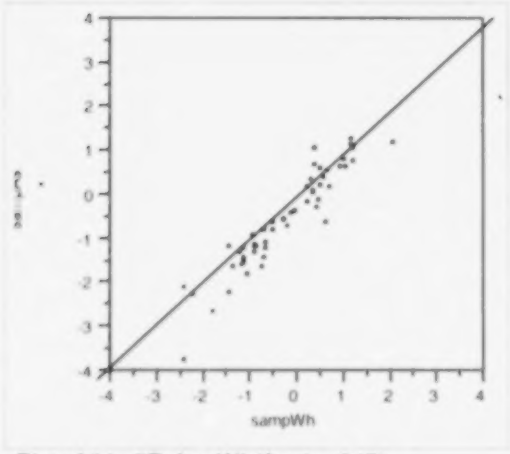


Plot of 'b'-2P for EPL/ESL (r=.934)

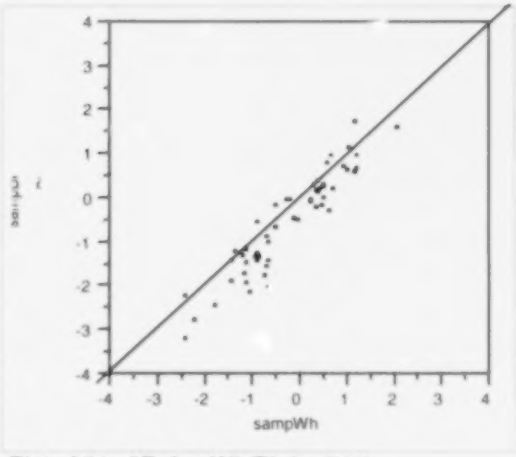
Verbal Reasoning (3P)



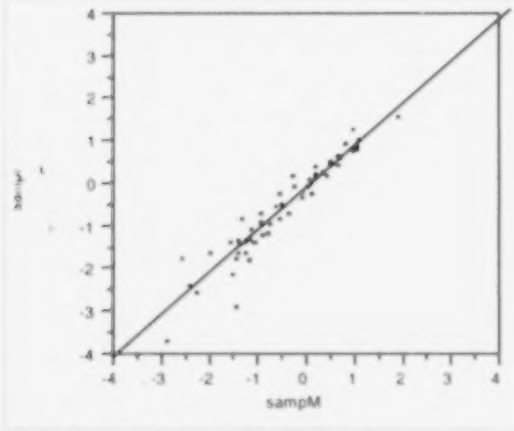
Plot of 'b' - 3P for Samples X/Y (r=.975)



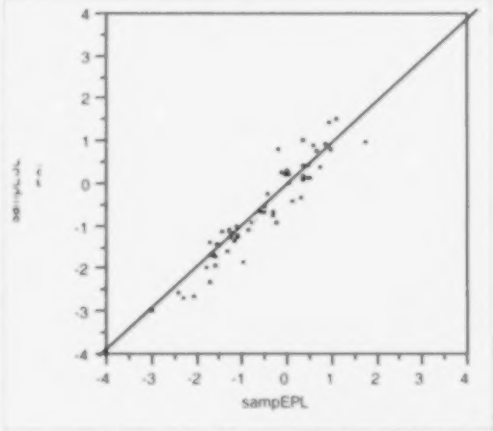
Plot of 'b' - 3P for Wh/As (r=.945)



Plot of 'b' - 3P for Wh/Bl (r=.944)



Plot of 'b' - 3P for M/F (r=.953)



Plot of 'b' - 3P for EPL/ESL (r=.949)

Appendix C

Table C1.

Mean ability estimates (and standard deviations) for the various subgroups using their own calibrations and sample X calibrations - across the 3 IRT models- PS.

Sample Groups	1P		2P		3P	
	Grp-calib.	X-calib.	Grp-calib.	X-calib.	Grp-calib.	X-calib.
Asian	0.429	0.369	0.396	0.351	0.399	0.339
	(0.985)	(0.965)	(0.976)	(0.950)	(0.890)	(0.925)
Black	-0.980	-0.881	-0.929	-0.871	-1.019	-0.927
	(0.653)	(0.743)	(0.601)	(0.714)	(0.823)	(0.796)
White	0.135	0.121	0.109	0.091	0.149	0.096
	(0.877)	(0.901)	(0.859)	(0.882)	(0.817)	(0.865)
Female	-0.223	-0.197	-0.232	-0.201	-0.198	-0.209
	(0.846)	(0.892)	(0.822)	(0.865)	(0.862)	(0.888)
Male	0.260	0.222	0.234	0.197	0.246	0.184
	(1.006)	(0.992)	(0.995)	(0.977)	(0.943)	(0.966)
EPL	0.012	0.014	-0.006	-0.007	0.022	-0.017
	(0.913)	(0.934)	(0.897)	(0.915)	(0.886)	(0.918)
ESL	0.087	0.067	0.075	0.056	0.074	0.027
	(1.101)	(1.070)	(1.081)	(1.049)	(1.077)	(1.072)

Table C2.

Mean ability estimates (and standard deviations) for the various subgroups using their own calibrations and sample X calibrations - across the 3 IRT models- VR

Sample Groups	1P		2P		3P	
	Grp-calib.	X-calib.	Grp-calib.	X-calib.	Grp-calib.	X-calib.
Asian	-0.080	-0.063	-0.109	-0.083	-0.078	-0.088
	(0.835)	0.881	(0.809)	0.855	(0.822)	0.863
Black	-0.818	-0.739	-0.802	-0.728	-0.830	-0.751
	(0.726)	0.820	(0.667)	0.767	(0.837)	0.808
White	0.255	0.232	0.219	0.205	0.241	0.198
	(0.788)	0.838	(0.803)	0.829	(0.735)	0.814
Female	-0.011	-0.011	-0.036	-0.025	-0.020	-0.034
	(0.938)	0.959	(0.917)	0.936	(0.922)	0.940
Male	0.029	0.032	0.003	0.009	0.022	0.000
	(0.898)	0.923	(0.887)	0.907	(0.875)	0.906
EPL	0.099	0.092	0.062	0.067	0.089	0.061
	(0.848)	0.884	(0.829)	0.861	(0.823)	0.859
ESL	-0.808	-0.714	-0.791	-0.701	-0.804	-0.728
	(0.780)	0.852	(0.717)	0.799	(0.841)	0.848

BEST COPY AVAILABLE

Appendix D

Table D1.
Mean ability estimates and standard deviations for all subgroups, from hard, easy and total test - PS.

Sample	Model 1P			Model 2P			Model 3P		
	Hard	Easy	Total	Hard	Easy	Total	Hard	Easy	Total
X	0.008	0.008	0.013	-0.005	-0.005	-0.001	-0.008	-0.007	-0.008
	(0.907)	0.868	0.963	0.908	0.853	0.943	0.883	0.867	(0.942)
Asian	0.340	0.269	0.369	0.341	0.244	0.351	0.334	0.245	0.339
	0.917	0.834	0.965	0.915	0.822	0.950	0.883	0.833	0.925
Black	-0.697	-0.827	-0.881	-0.704	-0.836	-0.871	-0.712	-0.855	-0.927
	0.643	0.825	0.743	0.634	0.791	0.714	0.617	0.812	0.796
White	0.072	0.139	0.121	0.048	0.109	0.091	0.047	0.109	0.096
	0.877	0.797	0.901	0.876	0.786	0.882	0.854	0.795	0.865
Female	-0.189	-0.139	-0.197	-0.187	-0.158	-0.201	-0.186	-0.163	-0.209
	0.812	0.885	0.892	0.810	0.860	0.865	0.797	0.876	0.888
Male	0.221	0.143	0.222	0.206	0.114	0.197	0.198	0.113	0.184
	0.932	0.868	0.992	0.934	0.857	0.977	0.906	0.869	0.966
EPL	0.003	0.024	0.014	-0.013	0.000	-0.007	-0.022	-0.003	-0.017
	0.871	0.872	0.934	0.877	0.853	0.915	0.859	0.866	0.918
ESL	0.149	-0.069	0.067	0.158	-0.098	0.056	0.149	-0.105	0.027
	0.967	0.954	1.070	0.959	0.944	1.049	0.941	0.961	1.072

Table D2.
Mean ability estimates and standard deviations for all subgroups, from hard, easy and total test - VR

Sample	Model 1P			Model 2P			Model 3P		
	Hard	Easy	Total	Hard	Easy	Total	Hard	Easy	Total
X	0.007	0.031	0.010	-0.011	0.000	-0.004	-0.015	-0.004	-0.013
	0.813	0.894	0.950	0.871	0.843	0.938	0.861	0.856	0.939
Asian	-0.063	-0.018	-0.063	-0.084	-0.054	-0.083	-0.086	-0.059	-0.088
	0.766	0.867	0.881	0.815	0.814	0.855	0.802	0.829	0.863
Black	-0.564	-0.641	-0.739	-0.622	-0.621	-0.728	-0.615	-0.639	-0.751
	0.672	0.913	0.820	0.709	0.823	0.767	0.684	0.848	0.808
White	0.179	0.234	0.232	0.177	0.188	0.205	0.167	0.188	0.198
	0.742	0.791	0.838	0.781	0.754	0.829	0.775	0.763	0.814
Female	-0.008	0.026	-0.011	-0.024	-0.001	-0.025	-0.025	-0.006	-0.034
	0.819	0.920	0.959	0.876	0.865	0.936	0.860	0.881	0.940
Male	0.014	0.037	0.032	-0.001	-0.007	0.009	-0.005	-0.011	0.000
	0.798	0.888	0.923	0.847	0.842	0.907	0.835	0.855	0.906
EPL	0.069	0.117	0.092	0.057	0.074	0.067	0.053	0.070	0.061
	0.775	0.846	0.884	0.821	0.797	0.861	0.808	0.810	0.859
ESL	-0.555	-0.640	-0.714	-0.599	-0.632	-0.701	-0.597	-0.650	-0.728
	0.696	0.934	0.852	0.736	0.844	0.799	0.715	0.866	0.848

BEST COPY AVAILABLE

Table E1.

Attenuated correlations between Proficiency estimates (theta) from odd/even and hard/easy tests for each sample across all models - 1P, 2P, 3P - Physical Sciences

	Odd-Even			Hard-Easy		
Sample	1 P	2 P	3 P	1 P	2 P	3 P
Random X	.80	.81	.81	.71	.73	.73
Asian	.78	.79	.79	.71	.73	.73
Black	.69	.71	.71	.56	.58	.62
White	.76	.78	.78	.68	.69	.70
Female	.76	.77	.78	.67	.68	.69
Male	.80	.81	.81	.73	.74	.75
EP	.77	.79	.79	.69	.71	.71
ESL	.83	.84	.84	.76	.78	.78

Table E2.

Disattenuated correlations between Proficiency estimates (theta) from odd/even and hard/easy tests for each sample across all models - 1P, 2P, 3P - Physical Sciences

	Odd-Even			Hard-Easy		
Sample	1 P	2 P	3 P	1 P	2 P	3 P
Random X	0.89	0.90	0.90	0.84	0.85	0.85
Asian	0.88	0.89	0.89	0.84	0.85	0.85
Black	0.82	0.84	0.84	0.73	0.74	0.77
White	0.87	0.88	0.88	0.82	0.82	0.83
Female	0.87	0.88	0.88	0.81	0.82	0.82
Male	0.89	0.90	0.90	0.85	0.86	0.86
EP	0.88	0.89	0.89	0.82	0.84	0.84
ESL	0.91	0.92	0.92	0.87	0.88	0.88

BEST COPY AVAILABLE

Appendix E

Table E3.

Attenuated correlations between Proficiency estimates (theta) from odd/even and hard/easy tests for each sample across all models - 1P, 2P, 3P - Verbal Reasoning.

	Odd-Even			Hard-Easy		
Sample	1 P	2 P	3 P	1 P	2 P	3 P
Random X	.78	.81	.81	.71	.73	.73
Asian	.75	.77	.77	.67	.68	.68
Black	.72	.74	.74	.63	.63	.64
White	.72	.74	.75	.64	.65	.65
Female	.80	.81	.81	.69	.70	.70
Male	.75	.77	.77	.70	.70	.71
EPL	.75	.77	.77	.66	.66	.66
ESL	.76	.77	.77	.62	.62	.63

Table E4.

Disattenuated correlations between Proficiency estimates (theta) from odd/even and hard/easy tests for each sample across all models - 1P, 2P, 3P - Verbal Reasoning.

	Odd-Even			Hard-Easy		
Sample	1 P	2 P	3 P	1 P	2 P	3 P
Random X	0.89	0.90	0.90	0.84	0.86	0.86
Asian	0.87	0.88	0.88	0.82	0.82	0.82
Black	0.85	0.86	0.86	0.79	0.79	0.80
White	0.85	0.86	0.87	0.80	0.80	0.80
Female	0.90	0.90	0.90	0.83	0.84	0.84
Male	0.87	0.88	0.88	0.84	0.84	0.84
EPL	0.87	0.88	0.88	0.81	0.81	0.81
ESL	0.87	0.88	0.88	0.78	0.78	0.79

BEST COPY AVAILABLE

Appendix F

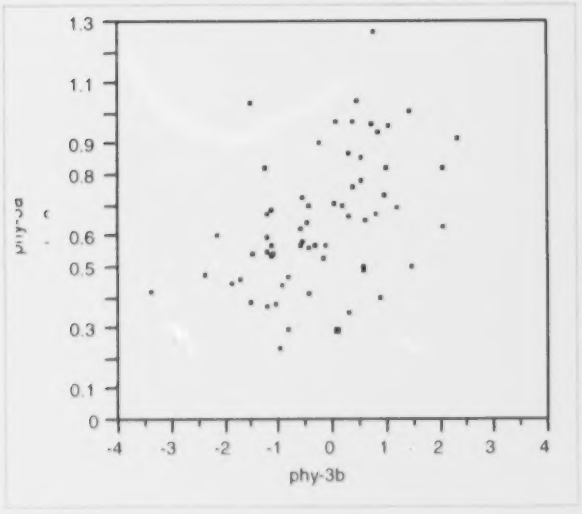


Figure F1. Plot of parameters a/b for 3P (Sample X - PS)

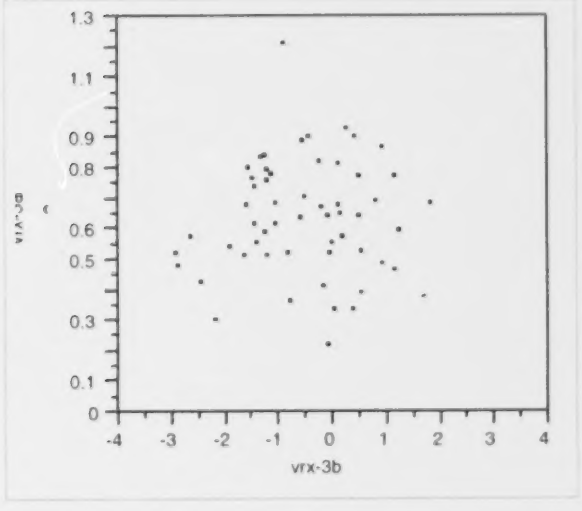


Figure F2. Plot of parameters a/b for 3P (Sample X - VR)

Appendix G

Table G1.

Correlations between theta estimates, for sample X, obtained using sample X calibrations and estimates obtained using calibrations from other subgroups - (PS).

Group	Model		
	1 P	2 P	3 P
Asian	.9991	.9981	.9916
Black	.9991	.9952	.9970
White	9991	.9982	.9979
Female	.9991	.9974	.9973
Male	.9991	.9985	.9983
EP	.9991	.9981	.9978
Esl	.9991	.9981	.9970

Table G2

Correlations between theta estimates for sample X obtained using sample X calibrations and estimates obtained using calibrations from other subgroups - (VR).

	Model		
Group	1 P	2 P	3 P
Asian	.9986	.9981	.9980
Black	.9986	.9960	.9948
White	.9986	.9975	.9966
Female	.9986	.9982	.9981
Male	.9986	.9975	.9973
EP	.9986	.9975	.9974
Esl	.9986	.9955	.9961

Appendix H

Table H1.

Mean ability estimates (standard deviations) for sample X using various subgroup calibrations - PS

	Mean ability estimates for Sample X (SD)		
Calibration Group	1 P	2 P	3 P
Sample X	0.013 (0.963)	-0.001 (0.943)	-0.008 (0.942)
Asian	0.013 (0.964)	-0.004 (0.932)	-0.010 (0.925)
Black	0.012 (0.959)	-0.003 (0.921)	-0.017 (0.922)
White	0.012 (0.964)	0.006 (0.932)	-0.009 (0.930)
Female	0.013 (0.964)	-0.019 (0.934)	-0.025 (0.932)
Male	0.012 (0.962)	0.009 (0.933)	0.009 (0.924)
EPL	0.012 (0.963)	0.001 (0.931)	-0.004 (0.924)
ESL	0.013 (0.960)	-0.005 (0.929)	-0.005 (0.922)

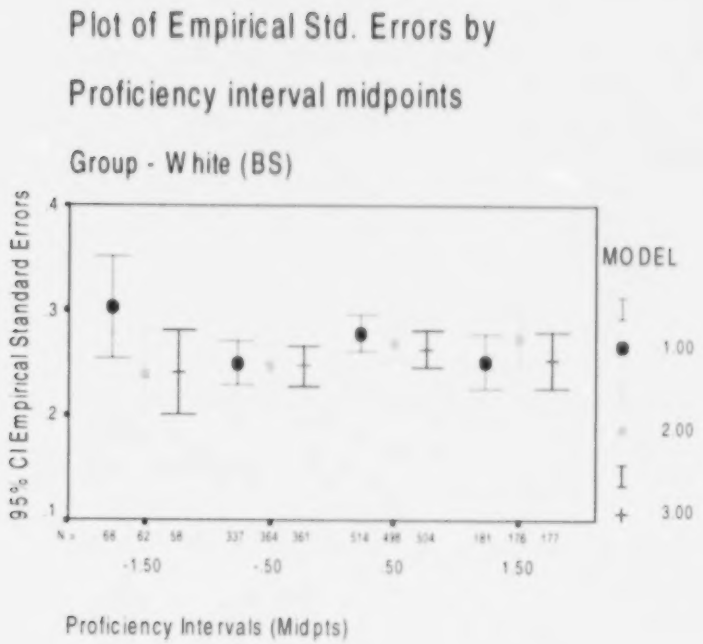
Table H2.

Mean ability estimates (standard deviations) for sample X using various subgroup calibrations - VR

	Mean ability estimates for Sample X (SD)		
Calibration Group	1 P	2 P	3 P
Sample X	0.010 (0.950)	-0.013 (0.915)	-0.013 (0.939)
Asian	0.010 (0.949)	-0.008 (0.933)	-0.017 (0.930)
Black	0.010 (0.947)	-0.015 (0.925)	-0.036 (0.926)
White	0.010 (0.951)	-0.001 (0.930)	-0.009 (0.910)
Female	0.010 (0.950)	-0.003 (0.934)	-0.013 (0.935)
Male	0.010 (0.950)	0.001 (0.931)	-0.009 (0.928)
EPL	0.010 (0.951)	-0.001 (0.935)	-0.008 (0.933)
ESL	0.010 (0.947)	-0.039 (0.916)	-0.046 (0.927)

Appendix I

Figure 11.



BEST COPY AVAILABLE

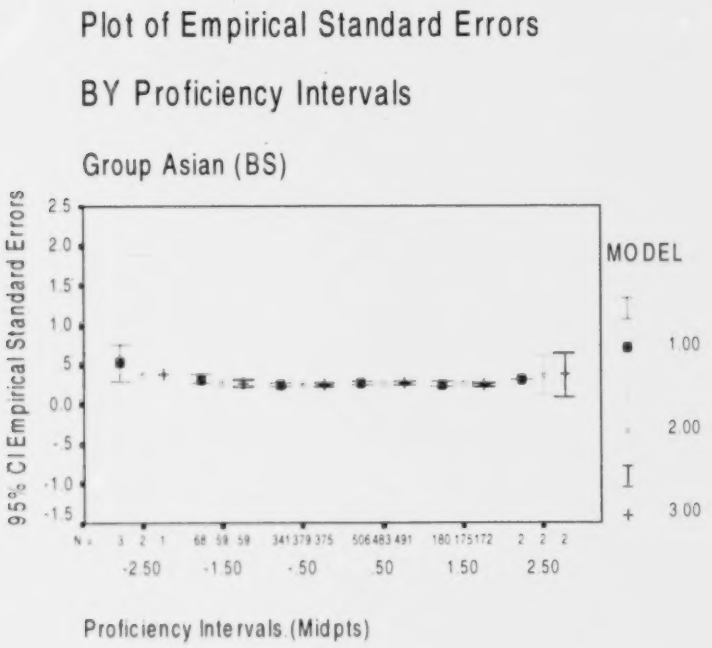


Figure I2.

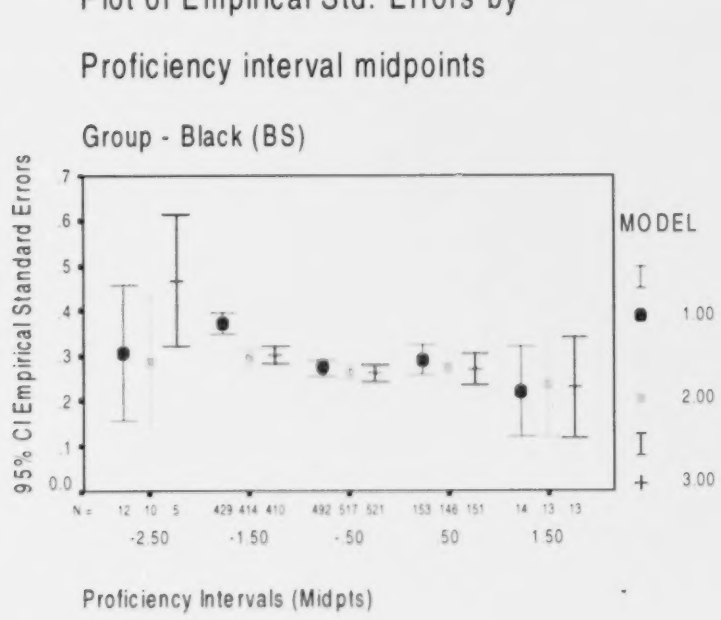


Figure I3.

BEST COPY AVAILABLE

Figure I4

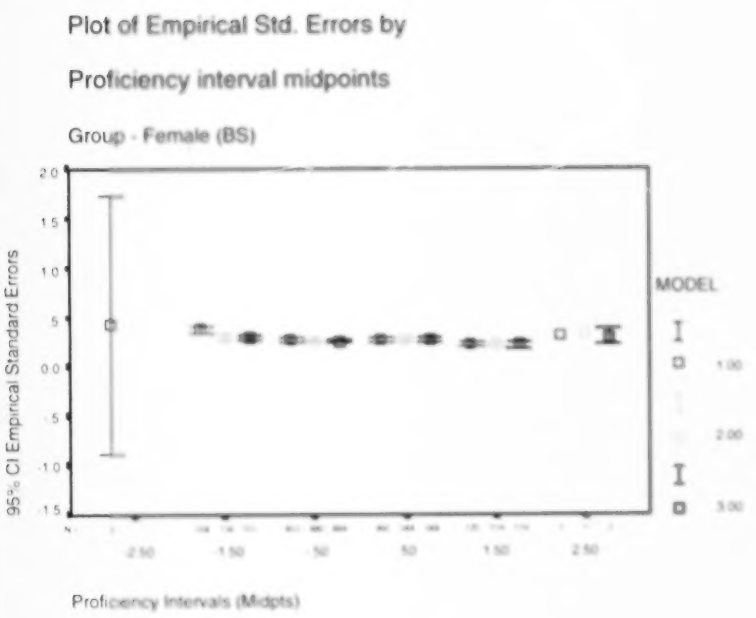
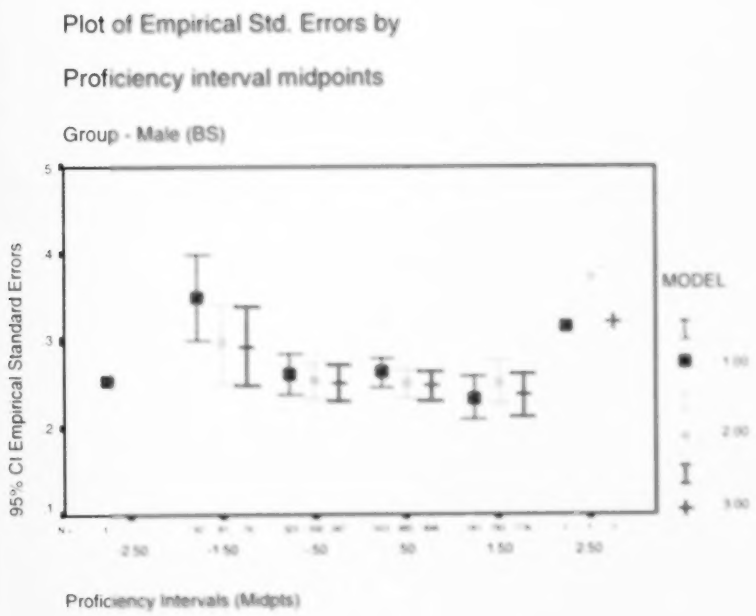


Figure I5.



BEST COPY AVAILABLE

Figure I6.

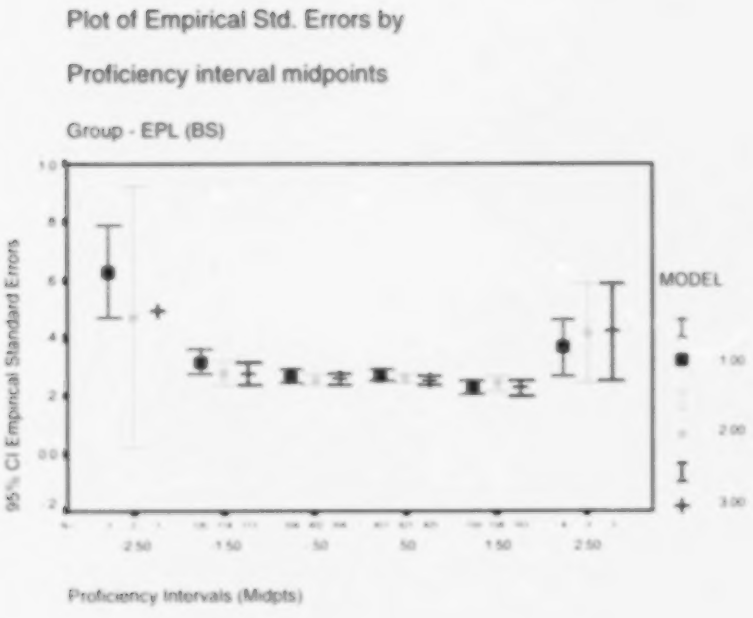
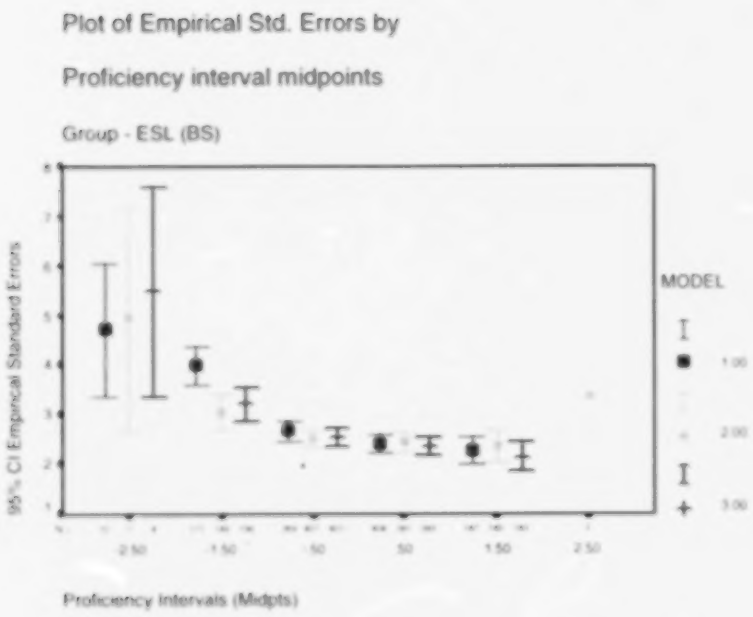


Figure I7.



BEST COPY AVAILABLE